Análise de Redes Sociais para Inferência de Similaridade entre Fornecedores de Governo

Tema: Relação governo e sociedade

Nº de páginas: 27

Folha de Rosto

Título do Trabalho: Análise de Redes Sociais para Inferência de Similaridade entre Fornecedores de Governo

Tema: Relação governo e sociedade

Resumo

A sociedade está desenvolvendo uma nova percepção a respeito da necessidade de transparência como um requisito de atuação na esfera pública, sejam Governos, empresas ou indivíduos. A ideia de que os processos de aquisição de Governo estão dominados pela corrupção de agentes públicos e associações entre fornecedores que agem sob má conduta econômica (ex: formação de cartéis) se tornou um senso comum, embora normalmente não haja evidência numérica que ajude a determinar o tamanho real do problema e, inclusive, explicite as estruturas virtuosas não contaminadas pelo comportamento fraudulento. Sociedades humanas, organizações e indivíduos são tipicamente organizados em redes. Existem em tais redes mecanismos indutivos que não são aleatórios e revelam modelos destas estruturas. Este trabalho aprofunda com rigor experimental a investigação sobre a relevância de modelos de indução de redes de fornecedores do Governo Federal brasileiro com base em seus padrões de venda, usando o arcabouço computacional de análise de redes sociais (SNA – *social network analysis*), ou análise de redes complexas. A complexidade é um atributo de sistemas – neste caso, redes – que exibem comportamento emergente e que não podem ser compreendidos através da análise isolada de suas partes. O Portal da Transparência do Governo Federal é um repositório de dados públicos. Entre esses dados públicos existem os gastos de Governo e cadastros de fornecedores inidôneos. Utilizamos estes dados para induzir uma rede de fornecedores com base em padrões de similaridade de venda para Governo e, depois, verificar a hipótese de que grupos de fornecedores conectados são discriminantes em relação a idoneidade, isto é, se estes grupos são formados quase totalmente por fornecedores idôneos ou inidôneos. Os resultados preliminares são expressivos e indicam que esta hipótese é verdadeira. O modelo de classificação possui acerto acima de 90% sobre a idoneidade de fornecedores, o que significa dizer que é quase sempre possível predizer, com base nos padrões de venda de um novo fornecedor, se ele é idôneo ou inidôneo. A estrutura do modelo e as confirmações experimentais reforçam as hipóteses de que existem associações entre fornecedores e de que agentes públicos desempenham um papel importante na legalidade das transações financeiras com estas empresas. Sistemas analíticos deste tipo, baseados em mineração de dados, podem ser de extrema utilidade para os órgãos

competentes de investigação no Brasil, pois apoiam a tomada de decisão no sentido de reduzir o espaço de busca por alvos e hipóteses mais prováveis de serem confirmadas. Para a sociedade, certamente representa um passo adiante na busca por mais transparência pública, que se concretiza não apenas pela publicação dos dados, mas também pela sua compreensão.

Palavras-chaves: transparência pública, análise de redes sociais, análise de redes complexas, mineração de dados, associação entre empresas, compras de Governo, agrupamento, classificação.

Sumário

1. Introdução

O Brasil está passando por uma crise de suas instituições. Escândalos corporativos envolvendo corrupção de agentes públicos são assuntos recorrentes na mídia nacional e internacional, fortalecendo ainda mais a percepção pela população de que os processos de aquisição do Governo estão dominados por associações entre fornecedores que agem sob má conduta econômica e fiscal.

A corrupção revela a existência de um deficit estrutural na governança pública, representando uma ameaça potencial para o sistema político como um todo, com impactos econômicos que já estão sendo sentidos por todos, sobretudo pelos mais pobres.

A imprevisibilidade das ações de corrupção que estão infiltradas nos processos formais e burocracia da máquina pública dificultam sua detecção, que cada vez mais pressupõe mecanismos inteligentes baseados não apenas na aplicação direta da legislação e regras gerais de auditoria, mas principalmente o reconhecimento de sua dinâmica de mudança e sua natureza baseada em formação de redes de relacionamento entre agentes públicos e privados. Tipicamente, a corrupção é contingente e se concretiza através de eventos pouco explícitos, de baixa visibilidade e inesperados, portanto difíceis de serem detectados, e, quando detectados, difíceis de serem explicados [1].

Os problemas também trazem uma oportunidade. A sociedade está desenvolvendo, cada vez com mais densidade, a capacidade de entender a transparência pública como um requisito para atuação social, o que se reflete nas manifestações sociais neste início de década. Iniciativas como o Portal da Transparência da Controladoria-Geral da União (CGU) [2], lançado no final de 2004 no âmbito do Governo Federal, bem como a Lei nº 12527 de novembro de 2011, conhecida como a Lei de Acesso à Informação (LAI) [3], válida para todas as esferas, constituem avanços importantes no sentido transparência.

Contudo, torna-se óbvio que a mera publicação de dados da administração pública por si só não realiza o máximo potencial da transparência. Para isso, é preciso que sociedade e Governo se organizem para trazer mais compreensão sobre os dados publicados, o que implica domínio sobre tecnologias de análise e mineração de dados.

Este trabalho propõe um método baseado em análise de redes complexas, ou como é mais conhecida, análise de redes sociais (SNA – *Social Network*

Analysis) [4], para detecção (inferência) de associações entre fornecedores de Governo com base em seus padrões de venda, extrapolando para um modelo de classificação sensível à idoneidade dos fornecedores. Os dados sobre gastos de Governo e idoneidade foram todos obtidos do Portal da Transparência da CGU.

A estrutura baseada em redes do modelo, bem como a relevância dos resultados experimentais, reforçam as hipóteses de que existem associações entre fornecedores e de que agentes públicos desempenham um papel importante na legalidade das transações financeiras. Ao mesmo tempo em que sistemas analíticos deste tipo são importantes para a sociedade no sentido de dar mais transparência pública, eles também se revelam como extremamente úteis para os órgãos de investigação e fiscalização no Brasil, pois apoiam a tomada de decisão no sentido de reduzir o espaço de busca por alvos e hipóteses mais prováveis de serem confirmadas. Em qualquer dos casos, a generalização da aplicação para outros cenários pode se revelar como oportunidade de negócio para o SERPRO, em particular no que diz respeito às funções de análise do Centro de Informação SERPRO (CIS) [5].

O documento está organizado da seguinte forma. A seção 2 apresenta as hipóteses de associação entre fornecedores de Governo consideradas no trabalho. A seção 3 traz alguns conceitos de análise de redes complexas úteis para o entendimento do modelo descrito na seção 4. A seção 5 descreve os experimentos e discute os resultados alcançados, demonstrando a validade do modelo. A seção 6 apresenta as conclusões e possíveis trabalhos futuros.

2. Associações entre Fornecedores de Governo

O estudo sobre a regularidade de transações entre empresas e Governo é um tema já discutido de longa data, porém recorrentes devido a sua atualidade. Stigler e Friedland criaram o termo “captura regulatória” em seus estudos sobre a inefetividade das atividades do Estado baseadas em regulação, demonstrando como é possível criar mecanismos que intentam burlar estes mecanismos de regulação e responsabilização [6].

No que diz respeito a compras de Governo, o processo regulatório está principalmente fundamentado na Lei nº 8666/1993 [7], que estabelece normas gerais sobre licitações e contratos de Governo, e na Lei nº 8137/1990 [8], que define crimes contra a ordem tributária, econômica e contra as relações de consumo. Os mecanismos que visam burlar estes processos de regulação estão

normalmente baseados em associações impróprias entre fornecedores privados e agentes de Governo (órgãos, gestores, funcionários, etc), e também associações entre fornecedores para estratégias de acerto de preço, ou cartéis.

O problema de inferir associações entre fornecedores e agentes de Governo e associações dos fornecedores entre si é derivado das seguintes perguntas/hipóteses:

1. Os fornecedores agem sozinhos? Existem associações entre eles?
2. Existem padrões de vendas que diferenciam fornecedores idôneos de inidôneos?
3. Se há associações entre os fornecedores, elas são hierárquicas?

Este trabalho aborda as duas primeiras perguntas, utilizando o arcabouço técnico de SNA e os dados de transações entre Governo e seus fornecedores.

3. Análise de Redes Complexas

Organizações humanas, de uma maneira geral, são normalmente organizadas em redes. Existem em tais redes mecanismos indutivos que não são aleatórios e revelam modelos destas estruturas [9]. Com a crescente disponibilidade de dados e a popularização das redes sociais virtuais, tais como Facebook e Twitter, o estudo sobre a formação de redes tem ganhado cada vez mais destaque.

A análise de redes complexas, ou SNA, investiga a estrutura e formação de redes sociais, humanas ou não (*e.g.* redes de documentos na Web), usando o arcabouço computacional da teoria dos grafos [4]. Enquanto grafos, os agentes da rede estão mapeados para nós, e seus relacionamentos para arestas. Relacionamentos são interpretados de acordo com o contexto da análise, podendo representar amizade (com arestas não direcionadas), um usuário seguindo outro (arestas direcionadas), ou qualquer outra relação, como ligações entre usuários com comunidades em comum no #você.serpro ou, a exemplo deste trabalho, similaridade entre fornecedores de Governo. A figura 1 ilustra um exemplo de representação em grafo de uma rede de amizades fictícia formada por 34 pessoas.

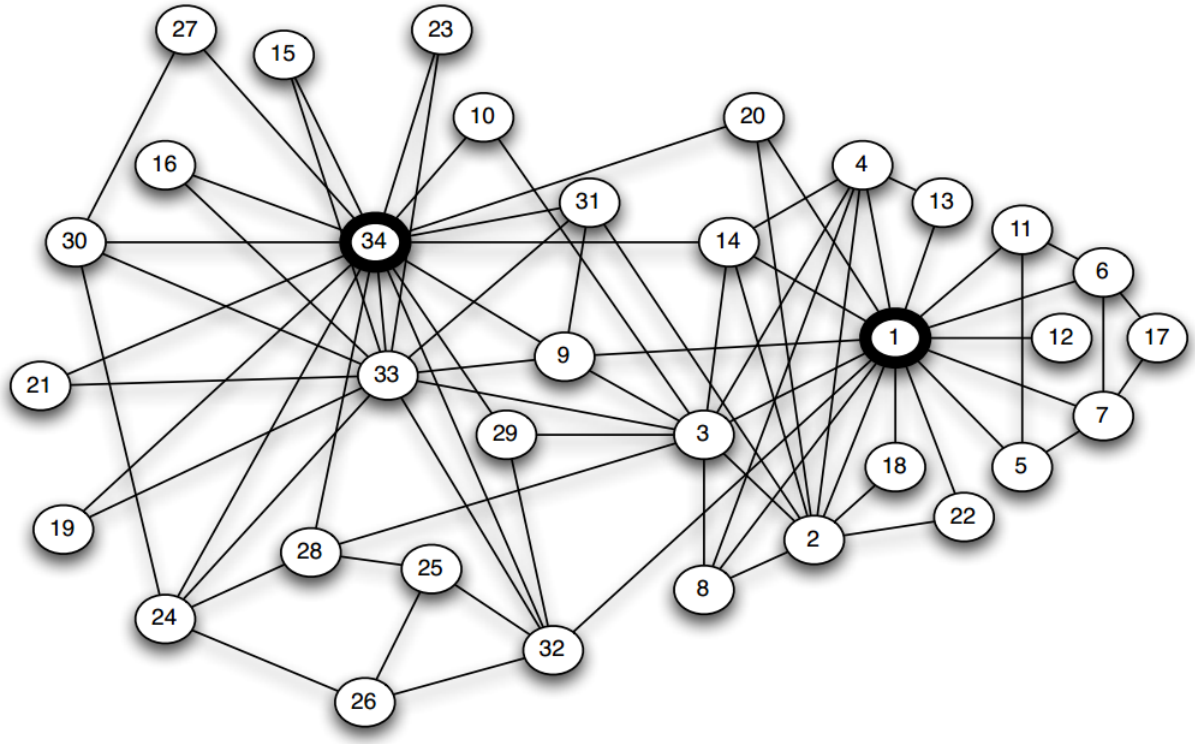

Figura 1. Rede de amizades fictícia de 34 pessoas. (Fonte: [4])

SNA emprega um grande conjunto de técnicas computacionais para análise de redes. Por exemplo, as técnicas de centralidade são métodos quantitativos importantes para detecção e comparação de nós e arestas influentes no grafo. Em destaque na figura 1 estão os influentes nós 1 e 34, ambos sendo os que possuem o maior número de conexões com outros nós (centralidade de grau).

Neste trabalho em particular foi utilizada uma métrica de centralidade de arestas baseada em *betweenness* (intermediação) [4]. O cálculo do *betweenness* das arestas identifica os relacionamentos mais importantes em uma rede no que diz respeito a quantidade de fluxo que elas carregam. A quantidade de fluxo que uma aresta carrega é diretamente proporcional à quantidade de vezes que esta aresta está presente nos menores caminhos entre os nós da rede. Isto significa que se uma aresta da rede com um alto valor relativo de *betweenness* fosse removida, os menores caminhos entre muitos pares de nós teriam que ser “re-roteados” em caminhos mais longos.

A remoção de arestas com alto *betweenness* deve produzir mais componentes conectados, ou *clusters* (conjuntos de nós conectados do grafo),

tendo sido bastante utilizada em SNA para identificação de comunidades. O algoritmo de Girvan-Newman, é um exemplo que se baseia nesse princípio [4]. Como será visto na próxima seção, o conceito de identificação de comunidades é aplicado para detecção de associações entre fornecedores com base em seus padrões de venda para Governo.

4. Modelo de Redes de Fornecedores de Governo

Um método baseado em SNA foi proposto para inferir grau de similaridade entre fornecedores a partir de suas afiliações no contexto de transações financeiras (gastos) envolvendo órgãos do Governo Federal. Afiliações são extraídas de transações classificadas como pagamento direto (de órgãos de Governo para fornecedores) e cartões corporativos (de pessoas autorizadas para fornecedores). O modelo final produzido, isto é, a rede inferida de fornecedores, tem o intuito de produzir uma aproximação da rede real de fornecedores, com arestas ponderadas pela probabilidade de comportamento similar.

Uma das motivações não técnicas do presente estudo está fortemente ligada à investigação de fraudes financeiras (*e.g.* apropriação indevida e lavagem de dinheiro) envolvendo recursos públicos. Nestes cenários, a similaridade entre agentes inidôneos (não confiáveis) é importante. Embora esta abordagem esteja relativamente longe de ser uma fonte de evidência (ou prova) de envolvimento de empresas em crimes financeiros, ela auxilia o processo de investigação ao fornecer como ponto de partida um universo de análise muito mais reduzido, potencialmente aumentando o foco e velocidade do processo, tomando como parâmetro a similaridade de comportamento com fornecedores já declarados e reconhecidamente considerados inidôneos.

Dados de gasto de Governo estão disponíveis no Portal da Transparência. Este trabalho está particularmente interessado nas seções de gastos diretos e transferências financeiras. Gastos diretos podem ser de dois tipos: pagamentos de órgãos de Governo para empresas; e Cartão de Pagamento do Governo Federal (CPGF). Em relação às transferências estão sendo considerados apenas os dados de Cartão de Pagamentos da Defesa Civil (CPDC), que faz uso dos recursos federais repassados pelo Ministério da Integração Nacional no contexto de convênios com estados e municípios, visando o apoio às ações de socorro. Fornecedores inidôneos também estão disponíveis no portal, o que inclui o Cadastro de Empresas Inidôneas e Suspensas (Ceis) e Cadastro de Entidades

Sem Fins Lucrativos Impedidas (Cepim).

O modelo final de associações entre fornecedores é produzido em dois passos, descritos nas subseções a seguir.

4.1. Rede de Afiliações dos Fornecedores

Os dados de gastos de Governo permitem extrair as seguintes conexões (ilustradas pela figura 2):

1. *Órgão de Governo e fornecedor* – conexão entre órgão de Governo (afiliação) e fornecedor (afiliado) ponderada pela quantidade total de dinheiro repassado pelo órgão à empresa fornecedora ao longo de um período de tempo;
2. *Indivíduo e fornecedor* – conexão entre portador de CPGF ou CPDC (afiliação) e fornecedor (afiliado) ponderada pela quantidade total de dinheiro repassado pelo indivíduo à empresa fornecedora ao longo de um período de tempo.

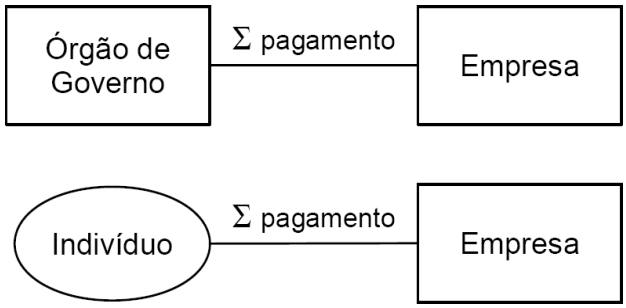

Figura 2. Afiliações de empresas (órgãos de Governo e indivíduos portadores de CPGF e CPDC).

O grafo de afiliações entre os fornecedores produzido não contém conexões entre as empresas, conforme ilustrado pela figura 3. A ideia é, a partir deste grafo, inferir um segundo grafo de conexões entre os fornecedores com arestas ponderadas pelo nível de similaridade entre eles. Essa similaridade é inferida a partir dos seus padrões de vendas, ou seja, a partir das afiliações e pesos das arestas neste primeiro grafo.

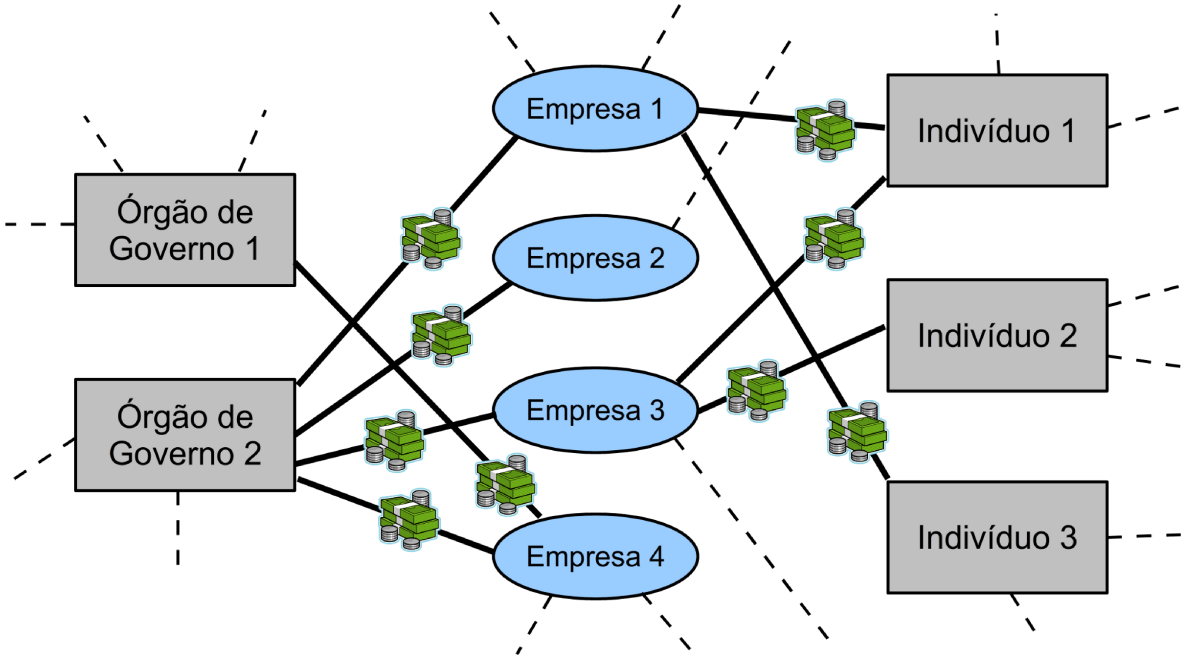

Figura 3. Grafo de afiliações de fornecedores.

4.2. Rede de Associações entre Fornecedores

Um segundo grafo, o de associações entre fornecedores, é gerado, tomando-se como ponto de partida a hipótese de que a rede de similaridades entre fornecedores pode ser inferida a partir de suas afiliações compartilhadas e pesos das suas conexões, e que empresas verdadeiramente similares estão no mesmo componente conectado[1] do grafo.

A similaridade entre um par de empresas *a* e *b* é obtido a partir de uma variação do coeficiente de similaridade de Jaccard [10]. O coeficiente original de Jaccard é útil para se obter o grau de similaridade entre dois conjuntos quaisquer, com base na quantidade de elementos em comum e nos tamanhos dos conjuntos originais (i.e. a métrica é robusta a conjuntos de tamanhos diferentes). Aplicado ao presente problema de similaridade entre empresas, ele seria definido em função dos conjuntos de vizinhos de ambas, N_a (vizinhos de *a*) e N_b (vizinhos de *b*), conforme a equação 1.

1 Um componente conectado(ou *cluster*) em um grafo não direcionado é formado por um conjunto de nós onde pelo menos um caminho entre qualquer par deles existe. Se não existe um caminho entre um par de nós A e B do grafo, então A e B estão em componentes conectados diferentes.

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

A variação do coeficiente tradicional de Jaccard é necessário para contemplar os valores de ponderação das arestas. Isto significa estender o conceito de similaridade para incluir também o de similaridade entre arestas, ou seja, tornar a similaridade entre as arestas de dois fornecedores para uma mesma afiliação proporcional à similaridade entre os seus valores. A equação 2 é responsável pelo cálculo de similaridade entre duas arestas (*a,c*) e (*b,c*)[2], onde $w_{(a,c)}$ e $w_{(b,c)}$ são os pesos de (*a,c*) e (*b,c*), respectivamente.

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

Quando $w_{(a,c)}$ e $w_{(b,c)}$ são iguais e positivos, a sub-expressão $|w_{(a,c)} - w_{(b,c)}|$ é igual a 0 (zero). Neste caso, a similaridade entre as arestas é 1. Por outro lado, quando $w_{(a,c)}$ e $w_{(b,c)}$ são diferentes e positivos, a similaridade é proporcional à diferença, inclusive considerando a magnitude dos valores, já que a diferença é normalizada pelo maior valor ($max(w_{(a,c)},w_{(b,c)})$).

O coeficiente de Jaccard alterado usa elementos das equações 1 e 2, e é representado pela equação 3. Em vez de considerar no numerador a quantidade de afiliações em comum, considera-se a soma das similaridades entre afiliações em comum.

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

O valor máximo de $\sum_{i \in N_a \cap N_b} s((a,i),(b,i))$ é $|N_a \cap N_b|$. Isto acontece quando os pesos de todas as conexões de *a* e *b* são iguais, situação em que a equação 3 se comporta como a equação 1. Usando a equação 3 é possível estabelecer a similaridade entre todas os fornecedores que possuem afiliações em comum, produzindo uma rede cujos nós são os fornecedores e as arestas o percentual de similaridade entre eles. A figura 4 ilustra esta nova rede.

2 Uma aresta (x,y) é uma aresta que conecta os nós x e y.

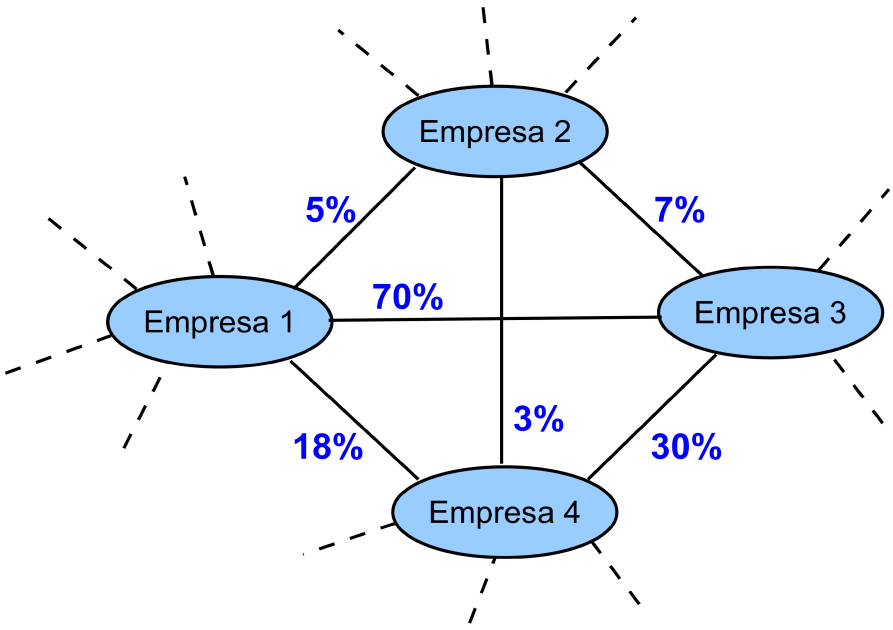

Figura 4. Rede de associações entre fornecedores.

Algumas conexões no grafo da figura 4 podem representar similaridades muito baixas. Pode-se estabelecer um limiar mínimo de similaridade entre fornecedores para remoção de arestas com valores muito baixos. Um limiar de 10% no grafo da figura 4 produziria o grafo da figura 5.

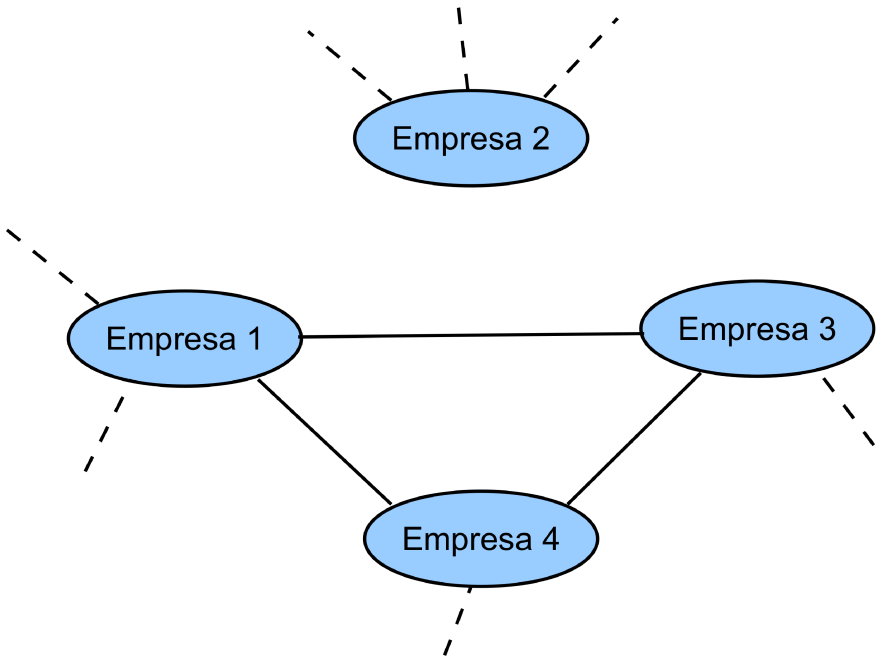

Figura 5. Grafo podado com limiar de corte de arestas de 10%.

Além da eliminação de arestas com valor abaixo de um limiar, também foi empregada em parte dos experimentos a eliminação de arestas com os maiores valores de *betweenness*. Neste caso, estabelece-se um percentual de arestas a serem removidas e elimina-se aquelas com maior *betweenness*.

Para validar a hipótese de similaridade entre empresas no mesmo componente conectado, foi investigado se os *clusters* formados são discriminantes em relação à idoneidade dos fornecedores. Isto significa dizer que é esperada a emergência de *clusters* quase totalmente formados por empresas idôneas ou quase totalmente formados por empresas inidôneas.

Após a formação da rede de associações entre as empresas, elas são rotuladas como idôneas ou inidôneas de acordo com o Ceis e Cepim. A figura 6 ilustra essa rotulação, exibindo os fornecedores idôneos na cor azul e os inidôneos na cor vermelha. Como demonstrado na figura, espera-se que os *clusters* sejam homogêneos em relação à idoneidade.

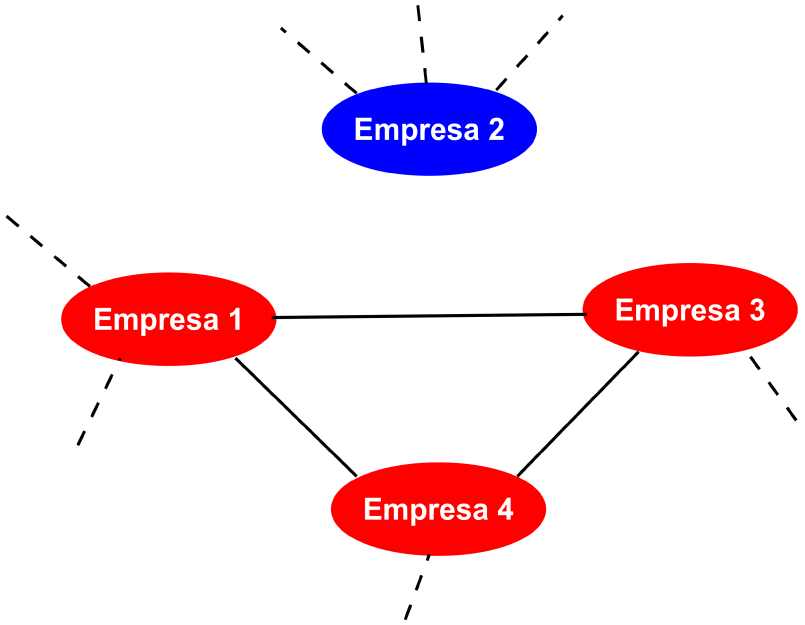

Figura 6. Rotulação de fornecedores quanto à idoneidade (idôneos são azuis e inidôneos são vermelho).

A rede de associações entre empresas rotuladas como idôneas ou inidôneas serve de modelo de classificação para outros fornecedores ainda não vistos. A inferência da idoneidade de um fornecedor desconhecido é ilustrada na figura 7. Embora se trate de uma inferência probabilística (*i.e.* que não garante

acerto em todos os casos), os elevados valores de acerto mostrados na próxima seção são evidências de que a hipótese de homogeneidade dos *clusters* é verdadeira.

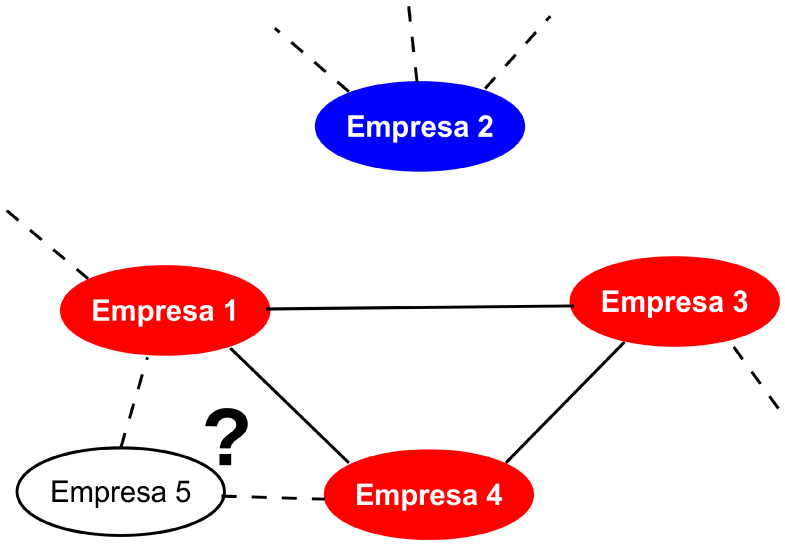

Figura 7. Inferência probabilística sobre a idoneidade de novos fornecedores.

5. Experimentos e Resultados

Há empresas idôneas e inidôneas. A hipótese a ser testada pelos experimentos é a de que fornecedores idôneos se comportam de maneira mais similar entre si, assim como os inidôneos. Em termos dos grafos gerados e de acordo com a mencionada hipótese, *clusters* (componentes conectados) tendem a ser homogêneos em relação à presença de companhias idôneas e inidôneas. Os experimentos estão organizados em dois grupos.

O primeiro grupo foi baseado nos seguintes dados:

- Gastos diretos coletados de janeiro de 2011 a março de 2014;
- CPDC de novembro de 2011 a abril de 2014;
- CPGF de janeiro de 2010 a março de 2014;
- Ceis de 09/05/2015;
- Cepim de 09/05/2014.

O segundo grupo de experimentos considerou os seguintes dados:

- Gastos diretos coletados de janeiro de 2011 a março de 2014[3];
- Ceis de 30/07/2015;
- Cepim de 30/07/2015.

Outra diferença entre os grupos de experimentos é que, no primeiro, a estratégia de formação da rede de associações entre fornecedores considerou apenas eliminação de arestas baseada apenas no limiar de similaridade entre empresas, enquanto que no segundo grupo explorou-se também a eliminação de arestas baseada em *betweenness*. No segundo grupo de experimentos também foi realizado o balanceamento de classes de idôneos e inidôneos, importante fator de justiça nos resultados de desempenho em classificação.

O principal objetivo da separação dos experimentos foi separar gastos de Governo de ordens de magnitude muito diferentes. Os gastos diretos são bem mais volumosos que os gastos de CPDC e CPGF. O primeiro grupo de experimentos deve ser considerado preliminar ao envolver todo tipo de gasto, enquanto que o segundo grupo mais rigoroso, pois se concentra nos gastos diretos e aplica método de remoção de arestas próprio para detecção de comunidades, isto é, apropriado para dar mais destaque e isolamento aos *clusters* formados.

Todos os experimentos foram realizados em uma máquina Dell R710, com 24 núcleos de processamento Intel Xeon X5650@2.67GHz e 24 GB de RAM. Os modelos foram desenvolvidos com apoio da ferramenta R [11] e da biblioteca Python Networkx [12]. Estas ferramentas estão entre as mais usadas em computação analítica e SNA.

5.1. Primeiro Grupo de Experimentos

As conexões entre as empresas indicam graus de similaridade (no intervalo entre 0,0 e 1,0) no que diz respeito aos serviços valorados prestados a órgãos de Governo e indivíduos autorizados. Conexões com peso 0 (zero) e companhias sem nenhuma conexão com as demais não foram incluídas no grafo.

O limiar mínimo de similaridade σ foi posto como parâmetro do modelo e foi usado na eliminação de arestas com valores inferiores ao definido, isto é, somente arestas com peso maior que σ foram mantidas no grafo. Neste primeiro grupo de experimentos, nenhuma métrica de centralidade, como *betweenness*

3 Não foram incluídos os dados de CPDC nem CPGF, ou seja, restou apenas órgãos de Governo como afiliações no grafo de afiliações.

de arestas, foram usadas para eliminação de arestas.

Foram experimentados quatro valores para σ: 0,25; 0,5; 0,75; e 1,0. A título de informação, uma conexão entre duas empresas com peso 1,0 significa que elas se comportam perfeitamente igual[4], enquanto que pesos cada vez mais próximos de 0,0 indicam menos similaridade entre as empresas. A tabela 1 exibe informações descritivas sobre os grafos formados para cada um dos valores de σ.

Tabela 1. Informações sobre os grafos do primeiro grupo de experimentos, para cada um dos valores de σ.

	$\sigma = 0,25$	$\sigma = 0,5$	$\sigma = 0,75$	$\sigma = 1,0$
# empresas	7377	6576	5773	2190
# conexões	409796	304773	215294	121164
Grau médio por nó	111,1	92,69	74,59	110,93
# *clusters*	524	716	780	386
# inidôneos	324	138	82	14

A figura 8 exibe as empresas e suas conexões para $\sigma = 1,0$. Cada *cluster* representa um grupo de fornecedores que possuem comportamento estritamente similar. É importante notar que, embora haja poucos inidôneos restantes, os *clusters* são muito homogêneos em relação a idoneidade. O caso de comportamento perfeitamente similar ($\sigma = 1,0$) torna-se bastante interessante em um cenário de investigação sobre associações ilegais entre empresas, uma vez que é extramente improvável que duas ou mais empresas, idealmente independentes, tenham exatamente o mesmo conjunto de vendas para Governo.

A figura 9 exibe a rede de empresas para $\sigma = 0,75$. Pode-se observar o aumento da agregação entre empresas sem afetar substancialmente a homogeneidade dos *clusters* em relação a idoneidade. Para $\sigma = 0,5$ (figura 10), assim como para $\sigma = 0,25$ (figura 11) o padrão de agregação de idôneos e inidôneos se mantém. Apenas a título de visualização, a rede para $\sigma = 0,05$ foi gerada (figura 12), mostrando que o modelo é extremamente discriminante em relação à idoneidade de fornecedores para Governo. Note o *cluster* central

4 Embora pareça extremamente improvável, ou virtualmente impossível, que duas ou mais empresas tenham exatamente o mesmo conjunto de vendas – isto é, vendas para os mesmos órgãos e indivíduos e extamente com os mesmos valores –, os resultados para σ=1 mostram que estes aglomerados de empresas existem!

formado em sua maioria por inidôneos (vermelhos) na figura 12, indicando que o modelo generativo capta o padrão de atuação em Governo dessas empresas e suas associações.

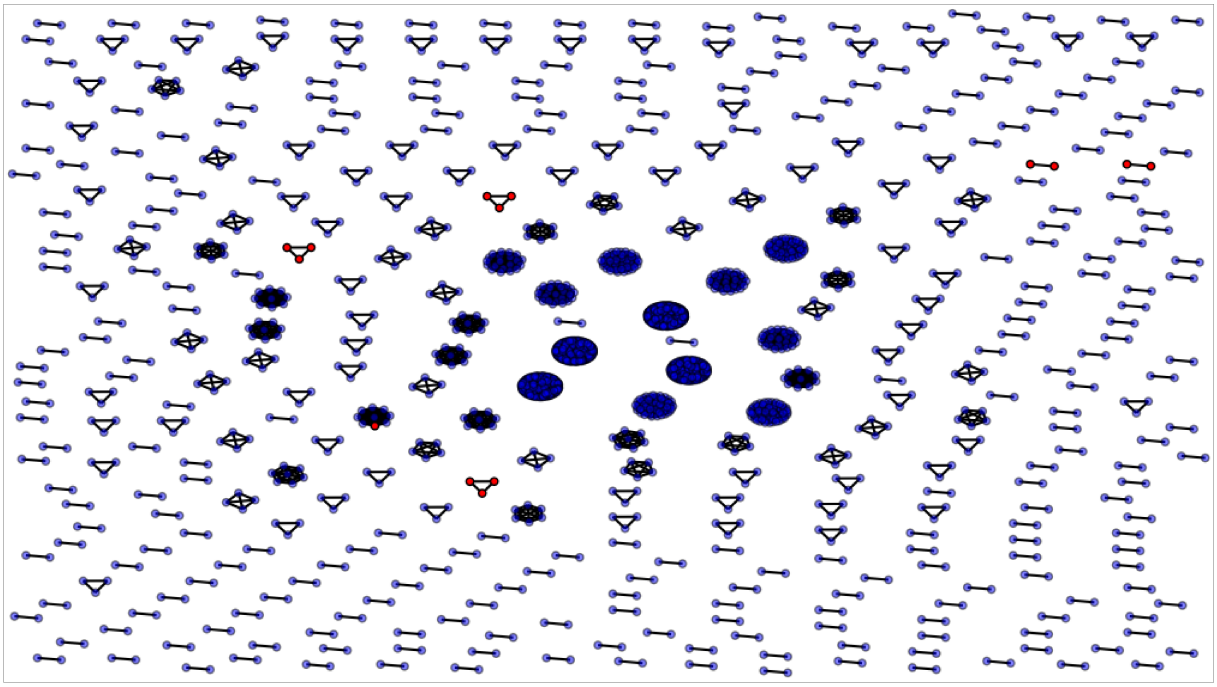

Figura 8. Rede de empresas idôneas (azuis) e inidôneas (vermelhas) para $\sigma = 1{,}0$.

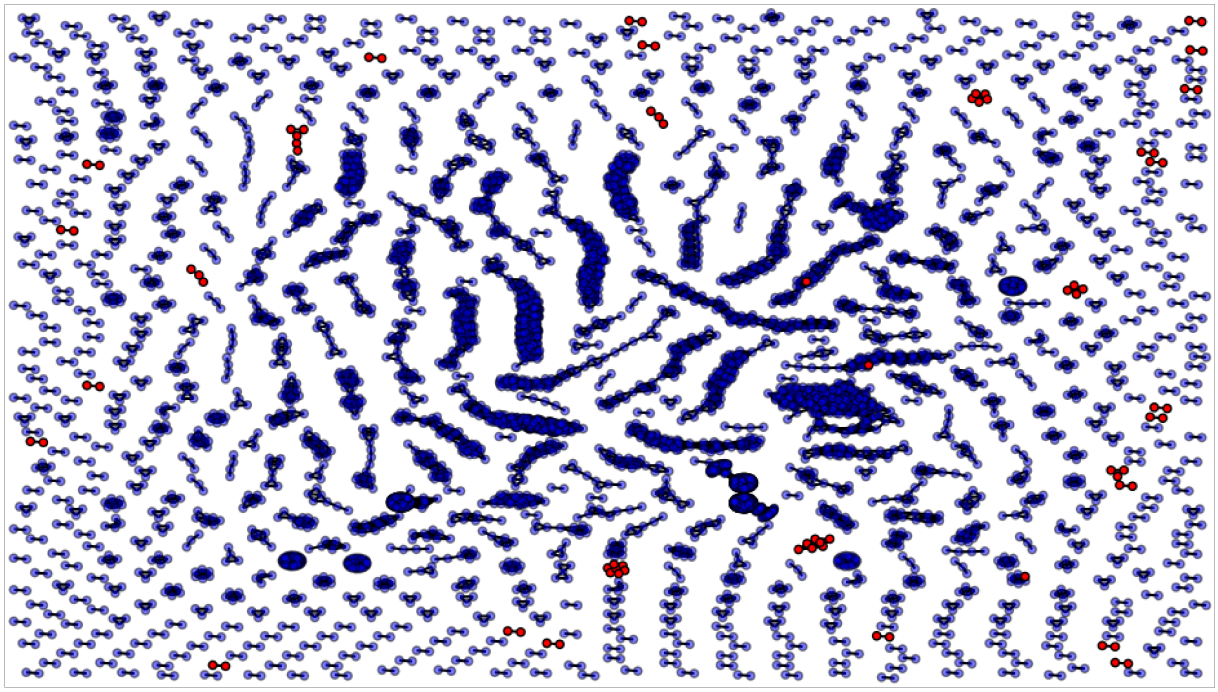

Figura 9. Rede de empresas idôneas (azuis) e inidôneas (vermelhas) para $\sigma = 0{,}75$.

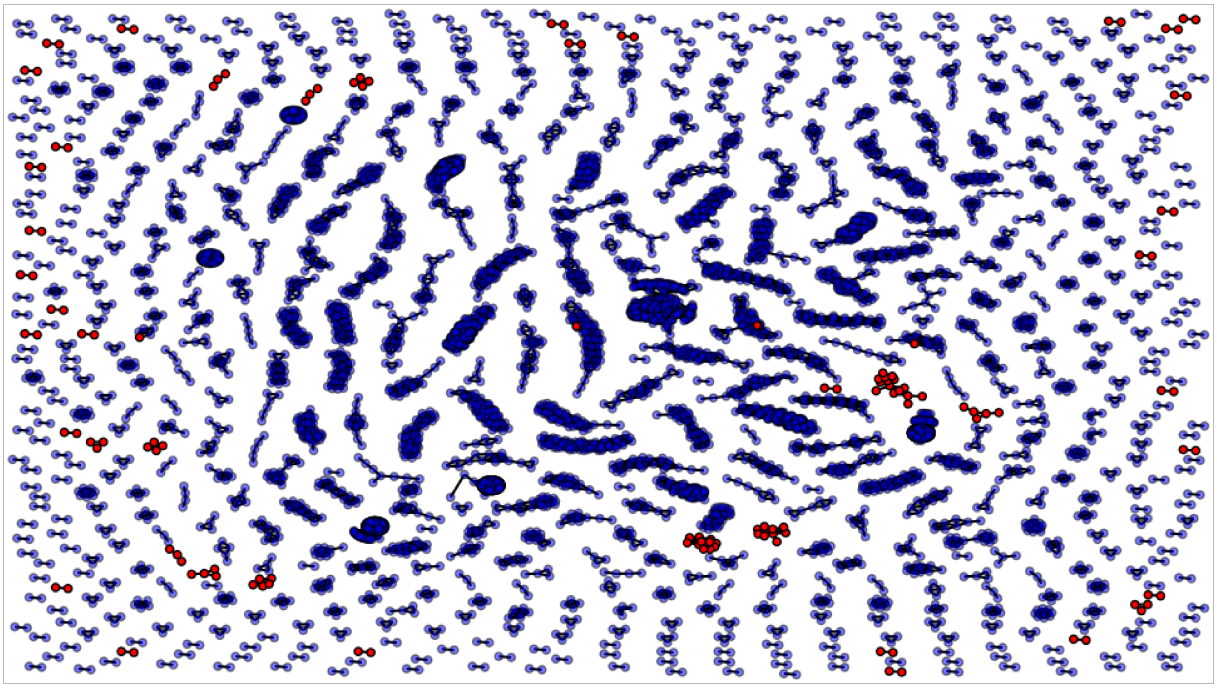

Figura 10. Rede de empresas idôneas (azuis) e inidôneas (vermelhas) para $\sigma = 0{,}5$.

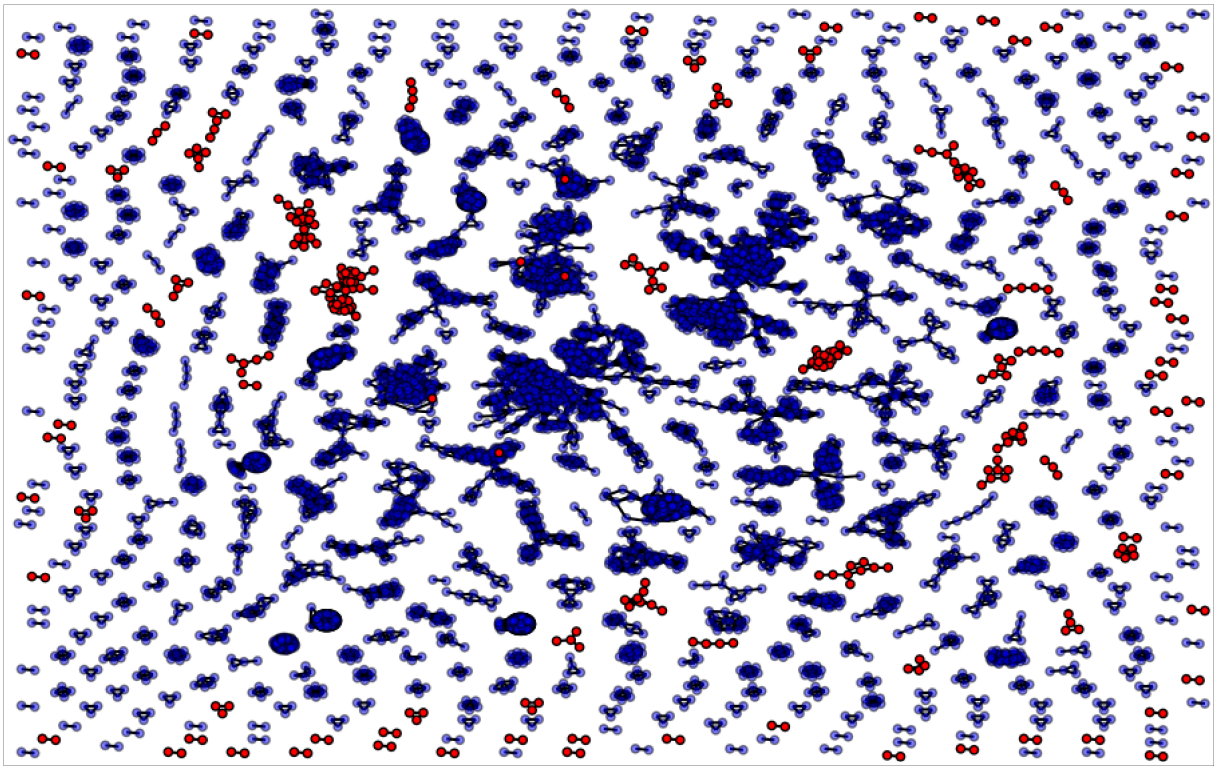

Figura 11. Rede de empresas idôneas (azuis) e inidôneas (vermelhas) para $\sigma = 0{,}25$.

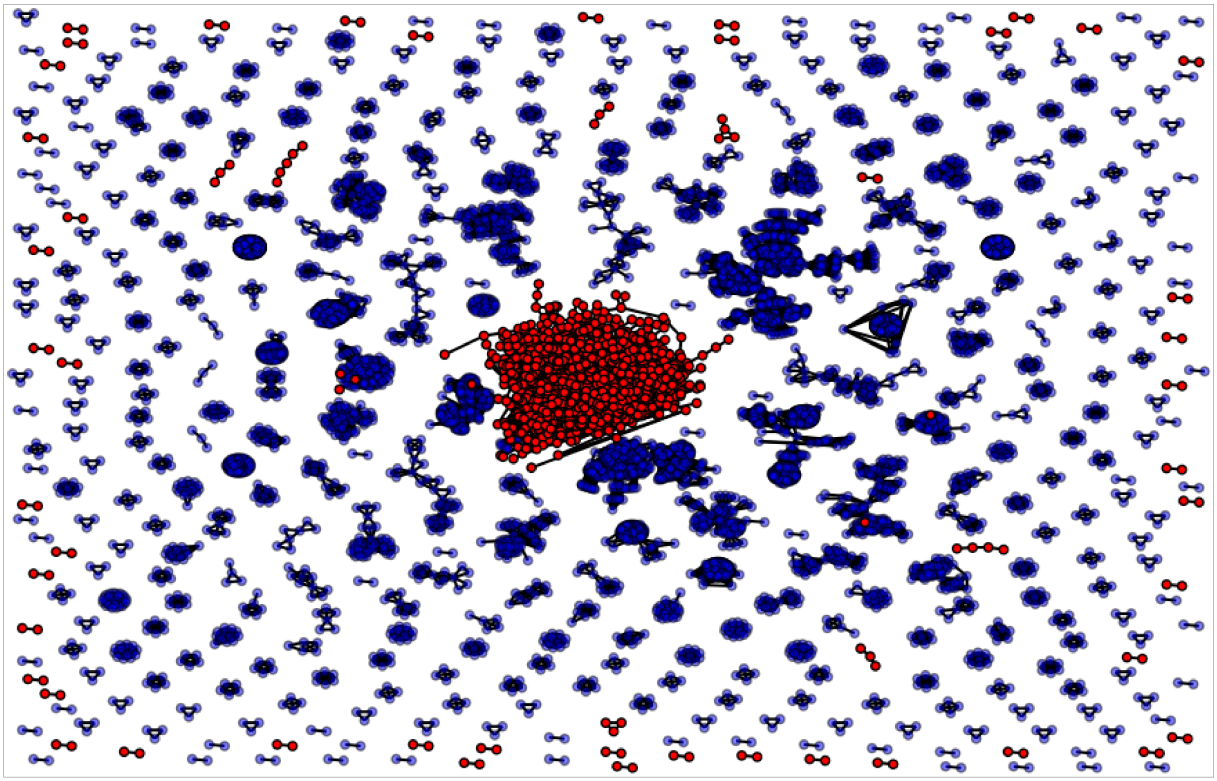

Figura 12. Rede de empresas idôneas (azuis) e inidôneas (vermelhas) para $\sigma = 0{,}05$.

Os resultados não foram apenas visuais. Foi medido o desempenho de classificação para os quatro grafos ($\sigma = 1{,}0$, $\sigma = 0{,}75$, $\sigma = 0{,}5$ e $\sigma = 0{,}25$) de acordo com a seguinte métrica. Seja C o conjunto de componentes conectados (*clusters*) do grafo. Cada componente conectado C^i é um conjunto de nós. C^i_{id} é subconjunto de C^i que contém apenas empresas idôneas, enquanto C^i_{inid} é o subconjunto das inidôneas. O desempenho de classificação do componente conectado C^i, denotado por π^i, é definido de acordo com a equação 4. A função de máximo (*max*) no numerador indica que o componente conectado representa a classe com mais elementos.

$$\pi^i = \frac{max\left(|C^i_{id}|, |C^i_{inid}|\right)}{|C^i|} \tag{4}$$

O desempenho global do grafo, denotado por $\prod$, é a média ponderada do desempenho de cada componente conectado. O fator de ponderação é o tamanho de cada *cluster*, de forma que erros de classificação em *clusters* maiores são mais penalizados do que em *clusters* menores. Este cálculo é

descrito pela equação 5.

$$\Pi = \frac{\sum_{i=1}^{|C|} |C^i| \pi^i}{\sum_{i=1}^{|C|} |C^i|} \quad (4)$$

Os resultados de desempenho para cada um dos valores de σ são mostrados na tabela 2. Todos as configurações possuem um acerto de classificação acima de 99,9%, indicando propriedade do modelo em diferenciar empresas idôneas de inidôneas. É impressionante perceber como a ciência das redes pode contribuir para o entendimento do comportamento emergente de desses agentes econômicos em cenários de Governo.

Tabela 2. Desempenho de classificação (idôneo/inidôneo) das redes de associações entre fornecedores para o primeiro grupo de experimentos, para cada um dos valores de σ.

	$\sigma = 0,25$	$\sigma = 0,5$	$\sigma = 0,75$	$\sigma = 1,0$
Π	99,932%	99,939%	99,948%	99,954%

Os resultados visuais e numéricos sobre desempenho de classificação do primeiro grupo de experimentos são animadores, apesar do não balanceamento de classes (números muito diferentes de idôneos e inidôneos), os dados estarem misturados (gastos diretos com CPDC e CPGF) e não ter sido aplicado nenhum corte baseado em *betweenness* de arestas. Sem ônus para os resultados já alcançados, este rigor foi levado para o segundo grupo de experimentos, descritos na próxima subseção.

5.2. Segundo Grupo de Experimentos

O segundo grupo de experimento herda todas as características do primeiro grupo, trazendo mais rigor experimental com as seguintes diferenças:

- Balanceamento de classes de idôneos e inidôneos. Como o número de idôneos é muito superior ao de inidôneos, para dar mais justiça à avaliação de desempenho de classificação decidiu-se reduzir os idôneos a uma amostra aleatória de tamanho igual ao conjunto de inidôneos;
- Inclusão de mais um valor para σ, 0,1;

- Considerou apenas gastos diretos de 2011 a 2014 e bases atualizadas (julho/2015) do Ceis e Cepim;
- Inclusão de procedimento de remoção de arestas com maiores valores de *betweenness*. Os percentuais de corte para cada valor de σ foi otimizado para produzir o menor erro de classificação.

A tabela 3 exibe o desempenho em classificação para o segundo grupo de experimentos, para cada um dos valores de σ. É interessante notar que a configuração com melhor desempenho é aquela com similaridade mínima de 10% entre os fornecedores. Exceto para σ=1.0, que se mostrou mais distante dos demais, o desempenho de classificação geral do modelo, considerando o maior rigor aplicado neste grupo de experimentos, é bastante aceitável. Para ficar mais claro, um acerto em classificação de 89,06% significa conseguir distinguir aproximadamente 9 em cada 10 fornecedores como idôneos ou inidôneos, isso com base em seus padrões de vendas para órgãos de Governo.

Tabela 3. Desempenho de classificação (idôneo/inidôneo) das redes de associações entre fornecedores para o segundo grupo de experimentos, para cada um dos valores de σ.

	σ = 0,1	σ = 0,25	σ = 0,5	σ = 0,75	σ = 1,0
Π	89,06%	86,78%	88,26%	88,62%	75%

As figuras de 13 a 17 exibem as redes de associações entre fornecedores para o segundo grupo de experimentos, para cada um dos valores de σ. Os resultados visuais, indicando poder de discriminação entre fornecedores idôneos e inidôneos, também são bastante promissores.

6. Conclusão e Trabalhos Futuros

Este trabalho apresenta uma investigação sobre a relevância de modelos de indução de redes de fornecedores do Governo Federal brasileiro com base em seus padrões de venda, usando o arcabouço computacional de SNA.

Os modelos produzidos levam em conta inicialmente a estrutura de afiliações com órgãos de Governo e indivíduos autorizados (portadores de CPGF e CPDC) e o quantitativo de repasses realizados ao longo de um período de tempo. Esses pagamentos para fornecedores revelam um padrão de vendas, o qual foi usado para produzir uma rede de associações entre empresas, com conexões

indicando nível de similaridade entre elas. O modelo final foi gerado após a remoção de parte das arestas, considerando similaridade mínima e, no segundo grupo de experimentos, *betweenness* de arestas. Dados do Ceis e Cepim foram usados para rotulação dos fornecedores como idôneos ou inidôneos.

Os resultados apresentados, tanto numéricos quanto visuais, mostram que os modelos apresentados geraram *clusters* que são discriminativos em relação à idoneidade das empresas, permitindo seu uso como modelo de classificação. Tanto nos experimentos mais preliminares (primeiro grupo) quanto nos mais rigorosos (segundo grupo) os resultados são bastante aceitáveis (acertos acima de 99% no primeiro grupo e 89% no segundo grupo).

O poder dicriminativo dos *clusters* reforça a hipótese de que existem associações entre fornecedores, com potencial aprofundamento para detecção de cartéis. Além disso, como o ponto de partida dos modelos são as afiliações com Governo, fortalece-se numericamente a hipótese de que agentes públicos (órgãos, empresas, funcionários, etc) desempenham um papel central na legalidade das transações entre fornecedores e Governo.

Dentre possíveis trabalhos futuros, destacamos os seguintes: (i) melhorar o modelo de rede com modelos econômicos mais completos, visando tornar ainda maior o desempenho em classificação; (ii) investigar como o trabalho pode ser adaptado e aplicado como uma ferramenta de ciência de dados na redução de espaço de busca por hipóteses plausíveis em instituições de investigação de crimes financeiros (*e.g.* COAF, PF, MPF, CGU, TCU), potenciais clientes do SERPRO; (iii) generalizar o trabalho de aplicação de SNA para outros domínios, não apenas para dados públicos (Portal da Transparência, dados.gov.br, etc), mas também dados do SERPRO e seus clientes, se for de interesse.

Uma perspectiva final ao observar as mudanças de modelos de negócio do SERPRO é a potencial exploração de ferramentas como a apresentada (ainda na qualidade de protótipo, mas passíveis de entrarem em produção) no modelo de análise como serviço, importante componente do Centro de Informações SERPRO (CIS). Especificamente quanto a esta aplicação, pode-se fortalecer o papel do SERPRO como agente importante para a sociedade, na busca por mais transparência pública, que se concretiza não apenas pela publicação dos dados, mas também pela sua compreensão.

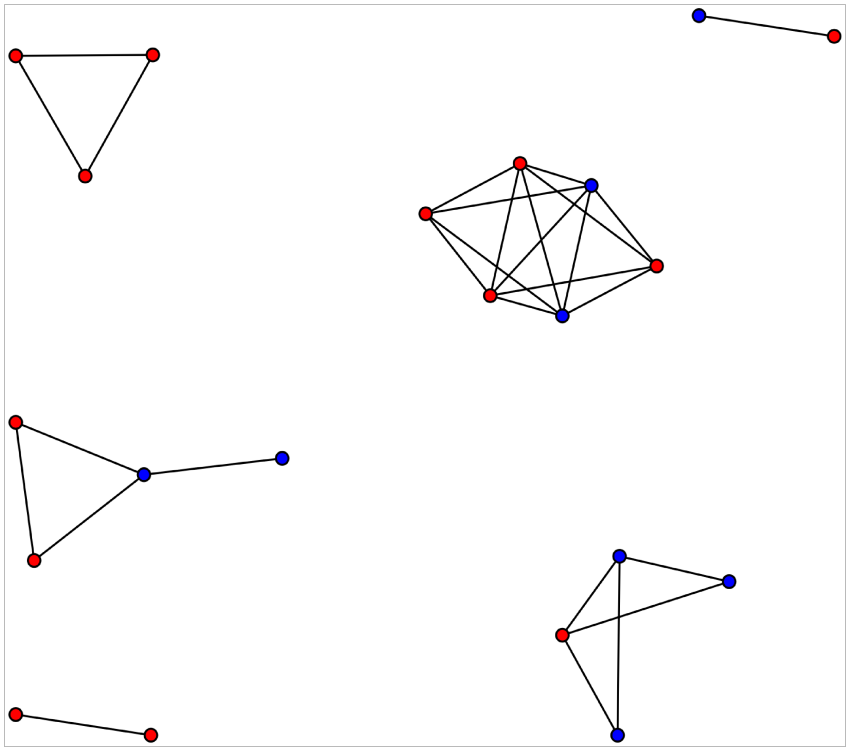

Figura 13. Rede de empresas idôneas (azuis) e inidôneas (vermelhas) para σ=1.0 e remoção de 13% das arestas de maior *betweenness*.

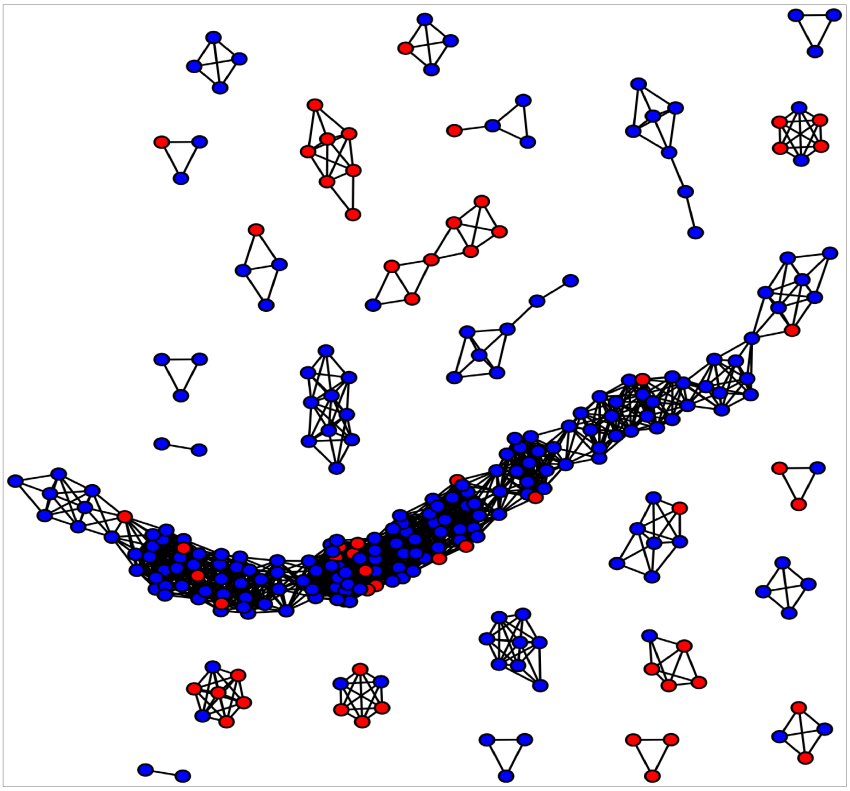

Figura 14. Rede de empresas idôneas (azuis) e inidôneas (vermelhas) para σ=0.75 sem remoção de arestas por *betweenness*.

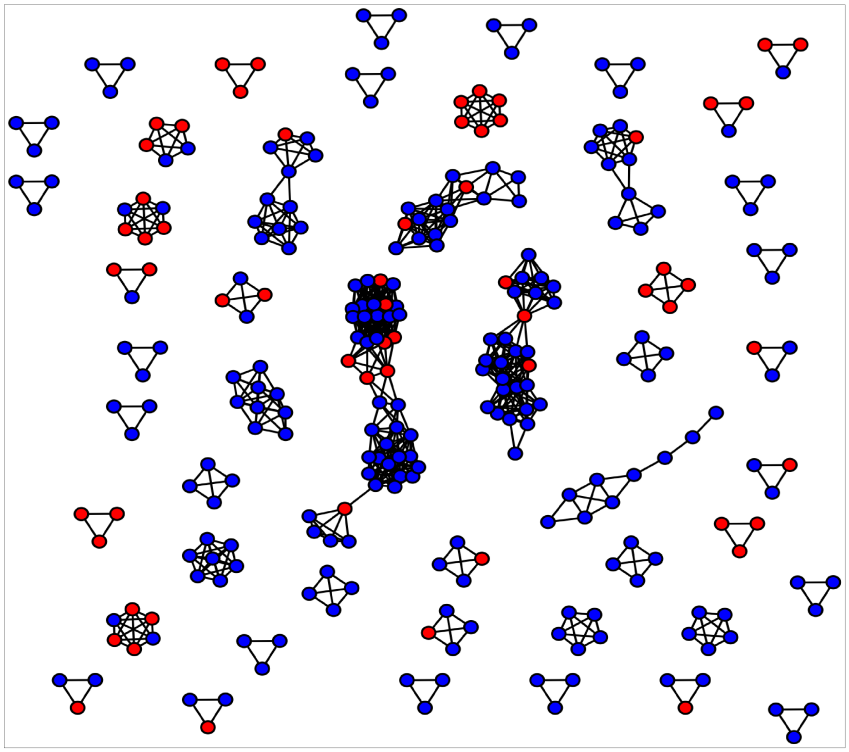

Figura 15. Rede de empresas idôneas (azuis) e inidôneas (vermelhas) para σ=0.5 e remoção de 80% das arestas de maior *betweenness*.

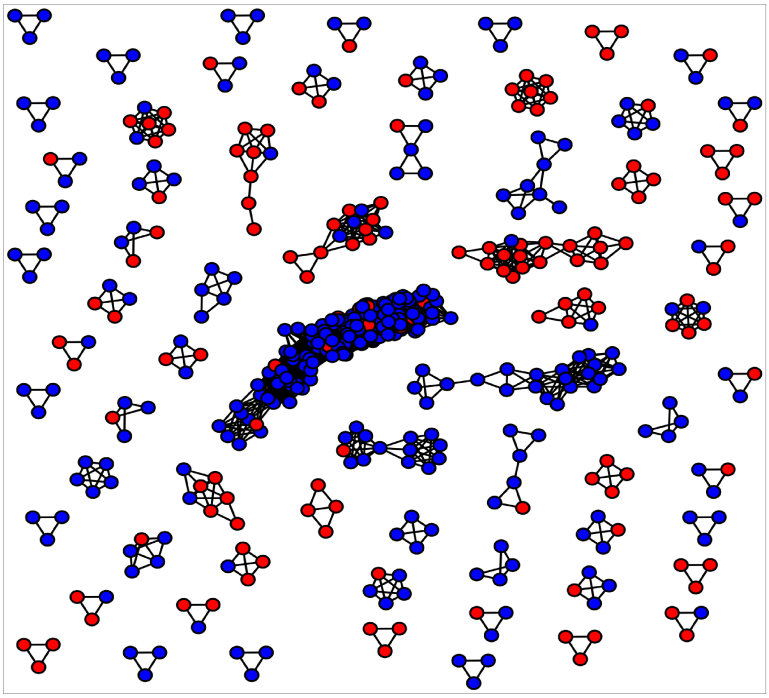

Figura 16. Rede de empresas idôneas (azuis) e inidôneas (vermelhas) para σ=0.25 e remoção de 40% das arestas de maior *betweenness*.

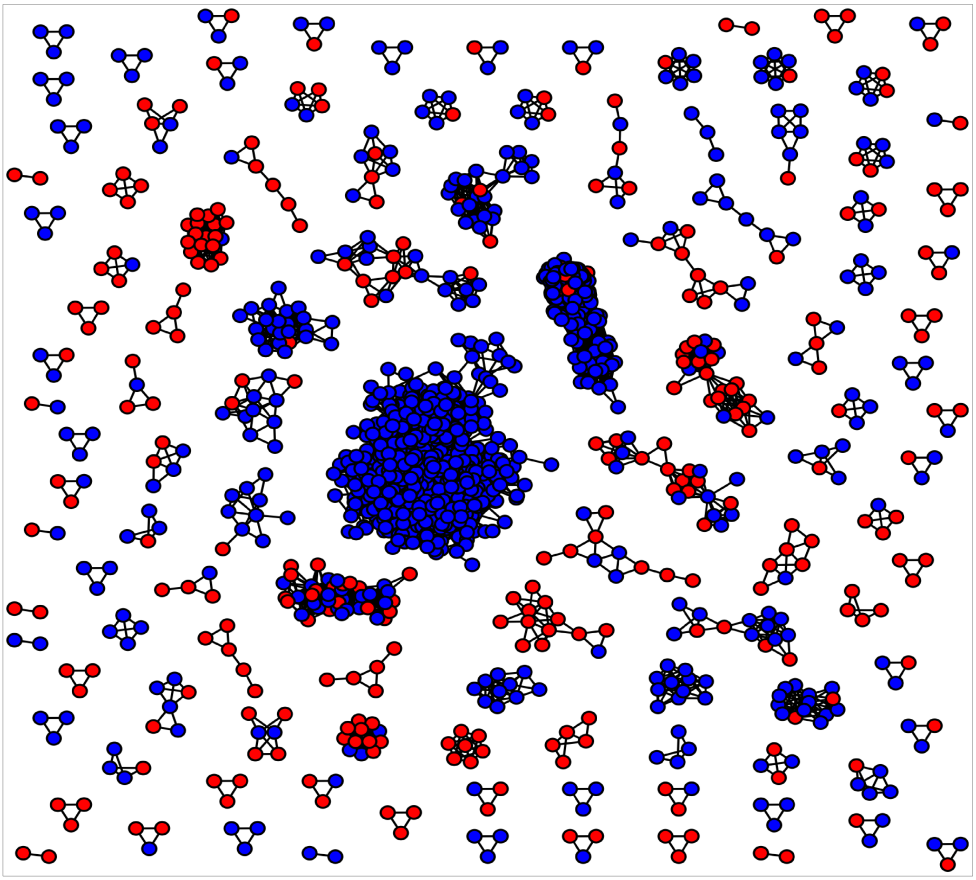

Figura 17. Rede de empresas idôneas (azuis) e inidôneas (vermelhas) para σ=0.01 e remoção de 35% das arestas de maior *betweenness*.

Referências

[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] CONTROLADORIA GERAL DA UNIÃO. Portal da Transparência, 2015. Disponível em <http://transparencia.gov.br/>. Acessado em 21 de agosto de 2015.

[3] PRESIDÊNCIA DA REPÚBLICA. LEI Nº 12.527, DE 18 DE NOVEMBRO DE 2011, 2011. Disponível em <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Acessado em 21 de agosto de 2015.

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

[5] SERPRO. Centro de Informações Serpro, 2015. Disponível em <http://cis.serpro .gov.br/>. Acessado em 21 de agosto de 2015.

[6] STIGLER, G; and FRIEDLAND, C. *What Can Regulators Regulate? The Case of Electricity.* In: Journal of Law & Economics, v. 5(1), 1962.

[7] PRESIDÊNCIA DA REPÚBLICA. LEI Nº 8.666, DE 21 DE JUNHO DE 1993, 1993. Disponível em <http://www.planalto.gov.br/ccivil_03/Leis/L8666cons.htm>. Acessado em 21 de agosto de 2015.

[8] PRESIDÊNCIA DA REPÚBLICA. LEI Nº 8.137, DE 27 DE DEZEMBRO DE 1990, 1990. Disponível em <http://www.planalto.gov.br/ccivil_03/leis/L8137.htm>. Acessado em 21 de agosto de 2015.

[9] RUTHS, J.; RUTHS, D. *Control Profiles of Complex Networks*. In: Science, v. 343, 2014, pp. 1373-1375.

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

[11] R. The R Project for Statistical Computing, 2015. Disponível em <https://www.r-project.org/>. Acessado em 21 de agosto de 2015.

[12] NETWORKX. High-productivity software for complex networks, 2015. Disponível em <https://networkx.github.io/>. Acessado em 21 de agosto de 2015.