Conselho Administrativo de Defesa Econômica
Departamento de Estudos Econômicos

Documento de Trabalho

Nº 004/2022

Metodologia para Identificação Automática de Grupos Econômicos em Análise Antitruste

José Gildo de Araújo Júnior
(Professor do Magistério Superior/Cade)

Brasília, agosto de 2022

CADE
Conselho Administrativo de Defesa Econômica

Ministério da Justiça e Segurança Pública
Conselho Administrativo de Defesa Econômica

Metodologia para Identificação Automática de Grupos Econômicos em Análise Antitruste

Departamento de Estudos Econômicos – DEE

SEPN 515 Conjunto D, Lote 4, Ed. Carlos Taurisano

Cep: 70770-504 – Brasília-DF

www.gov.br/cade

ISSN 2764-1031

Esse documento foi produzido pelo Departamento de Estudos Econômicos do Conselho Administrativo de Defesa Econômica.

José Gildo de Araújo Júnior

(Professor do Magistério Superior/Cade)

Sumário Executivo

Como inferir que empresas do mesmo setor devem ser vistas como parceiras ao invés de concorrentes? Como medir os impactos dessas parcerias na análise dos mercados relevantes? Como aplicar multas e penalidades razoáveis sem conhecer precisamente o grupo econômico no qual as empresas estão inseridas e foram beneficiadas do ilícito? Neste trabalho, é proposta uma metodologia para a construção automática de grupos econômicos por meio da base de dados pública da Receita Federal. Sua relevância perpassa os corredores da análise concorrencial, podendo ser útil em vários setores financeiros como bancos e corretoras. Especificamente para o Cade, a construção automática de grupos econômicos permite minimizar o tempo gasto com requisições e com a construção manual de grupos econômicos, na verificação e validação das informações apresentadas pelas empresas ao submeterem os formulário de notificação, auxiliar na análise de mercados relevantes e, por fim, nas questões que envolvem a dosimetria de aplicação de multas e penalidades.

Palavras-chaves: Grupos econômicos; Automatização; Benchmark.

LISTA DE ILUSTRAÇÕES

LISTA DE CÓDIGOS

LISTA DE ABREVIATURAS E SIGLAS

AC	Ato de Concentração
BFS	Breadth-First Search
Cade	Conselho Administrativo de Defesa Econômica
DFS	Depth-First Search
HHI	Herfindal-Hirschman Index
QSA	Quadro de Sócios e Administradores
SGBD	Sistema de Gerenciamento de Banco de Dados
SQL	Structured Query Language

LISTA DE SÍMBOLOS

$\sim$	Aproximadamente
$\in$	Pertence
$\emptyset$	Conjunto vazio
$\neq$	Diferente
$\cap$	Intersecção

SUMÁRIO

1. INTRODUÇÃO

Ao darem entrada em um processo de Ato de Concentração (AC) junto ao Conselho Administrativo de Defesa Econômica (Cade), as empresas interessadas informam, via formulário de procedimentos ordinários, questões referentes aos grupos econômicos que fazem parte. Em determinado ponto do processo essas informações são contrastadas para saber se alguma relação societária relevante deixou de ser informada, de modo a não comprometer análises e medições que emergem dessa informação, como por exemplo, o cálculo do índice de Herfindal Hirschman (HHI), ou em outros casos, a aplicação de multas e penalidades que são aplicadas proporcionalmente ao grupo econômico no ramo de atividade empresarial em que ocorreu a infração. Geralmente, esta verificação é realizada por meio da construção manual de grupos econômicos. Por si só essa construção manual cria inúmeros inconvenientes. Inicialmente, é importante enfatizar que a análise de um AC possui prazo para ser concluída. A depender do cenário em voga, a construção manual de grupos econômicos pode consumir muito tempo para ser concluída. Seja pela robusta rede de relações societárias que uma das empresas possui, seja pela necessidade de notificação de órgãos e empresas em busca de esclarecimentos, seja pelo tempo necessário em fazer-se cumprir o rito burocrático dessas atividades, o fato é que quanto mais tempo for consumido nesta atividade, menos tempo restará para atividades ainda mais complexas, como, por exemplo, a construção e análise de mercado relevante, simulação de cenários críticos, reflexão sobre os resultados e construção de pareceres.

De modo a mitigar os problemas descritos, este trabalho apresenta uma pro- posta de construção automática de grupos econômicos baseado nas relações societárias presentes nos dados públicos disponibilizados pela Receita Federal. Esta abordagem apresentou 71% de precisão e 48% de revocação[1] média em relação às informações apresentadas pelas empresas. Verificou-se posteriormente que todos os falsos-positivos encontrados, apesar de não informados ao Cade, existiam dentro da base da receita.

São contribuições desse trabalho:

- Concepção e construção de um algoritmo para construção automática de grupos econômicos;
- Construção de um benchmark de grupos econômicos útil para verificar, validar e compararquaisquer iniciativas de propósito semelhante;

[1]As métricas de precisão e revocação são apresentadas em detalhes na seção 4.2.5 deste trabalho.

- Inserção da base de dados pública da Receita Federal em um Sistema de Gerenciamento de Banco de Dados (SGBD) dentro do ambiente interno do Cade, possibilitando consultas complexas à empresas, sócios de empresas e grupos econômicos para os anos de 2020 e 2021;
- Definição matemática do problema de construção de grupos econômicos no contexto do Cade.

2. DEFINIÇÃO DO PROBLEMA

Neste ponto torna-se importante esclarecer a subjetividade do conceito de grupo econômico e como ele é compreendido em distintos campos de estudo. A seguir são apresentadas definições de grupos econômicos à luz de diferentes saberes.

2.1 Visão Jurídica

A Resolução Cade nº 33/2022 estabelece uma definição sobre grupo econômico da seguinte forma:

RESOLUÇÃO Nº 33, DE 14 DE ABRIL DE 2022 disciplina a notificação dos atos de que trata o artigo 88 da Lei nº 12.529, de 30 de novembro de 2011, prevê procedimento sumário de análise de atos de concentração e consolida as Resoluções nº 02/2012, 09/2014 e 16/2016. Art. 4º Entende-se como partes da operação as entidades diretamente envolvidas no negócio jurídico sendo notificado e os respectivos grupos econômicos. §1º Considera-se grupo econômico, para fins de cálculo dos faturamentos constantes do art. 88 da Lei nº 12.529/2011, cumulativamente: I - as empresas que estejam sob controle comum, interno ou externo; e II - as empresas nas quais qualquer das empresas do inciso I seja titular, direta ou indiretamente, de pelo menos 20% (vinte por cento) do capital social ou votante. §2º No caso dos fundos de investimento, são considerados integrantes do mesmo grupo econô- mico para fins de cálculo do faturamento de que trata este artigo, cumulativamente: I - O grupo econômico de cada cotista que detenha direta ou indiretamente participação igual ou superior a 50% das cotas do fundo envolvido na operação via participação individual ou por meio de qualquer tipo de acordo de cotistas; e II - As empresas controladas pelo fundo envolvido na operação e as empresas nas quais o referido fundo detenha direta ou indire- tamente participação igual ou superior a 20% (vinte por cento) do capital social ou votante. §3º A definição de grupo econômico deste artigo aplica-se apenas para fins de cálculo do faturamento com vistas à determinação do atendimento dos critérios objetivos fixados no artigo 88 da Lei nº 12.529/2011, e não vincula decisões do Cade com relação à solicitação de

informações e à análise de mérito dos casos concretos.

2.2 Visão Econômica

Segundo (ALMEIDA, 2007) a concorrência econômica corresponde à atuação independente entre vendedores de bens ou serviços frente à consumidores tendo por objetivo maximizar o lucro. Em um ambiente concorrencial, os diferentes vendedores desenvolvem suas estratégias e duelam entre si para ampliarem a quantidade de clientes e consequentemente aumentarem sua lucratividade.

Dessa forma, um grupo econômico pode ser definido como um conjunto de empresas que apesar de possuírem distintos cadastros de pessoas jurídicas, apresentam participação societária em comum, e assim, compartilham dos mesmos interesses econômicos tornando-se, para tanto, inexistentes, ou então, menos vigorosos os efeitos da concorrência sobre elas. Por exemplo, se a empresa **A** possui sociedade com a empresa **B**, a empresa **A** não pode mais ser analisada como concorrente da empresa **B** naquele ramo de atividade em que a sociedade foi constituída, uma vez que compartilham dos mesmos interesses. Sendo assim, para fins de análise concorrencial deve-se passar a enxergar o grupo econômico **A–B** uma vez que ambas as empresas participantes de uma estrutura societária fazem frente a concorrência juntas, subordinando-se aos mesmos interesses do grupo a que pertencem.

2.3 Definição do Problema

A seguir é apresentada a ideia central, seguida da formalização matemática do problema a ser mitigado.

2.3.1 Ideia Central

Dado um identificador de uma empresa, o cnpj por exemplo, o objetivo do problema sendo analisado é retornar o conjunto de empresas que fazem parte do grupo econômico em que a empresa sendo analisada está inserida. Em grande medida, esta informação é semelhante às informações declaradas pela empresa em questão no item II.5 do formulário de notificação dos atos de concentração submetidos ao Cade.

2.3.2 Formalização Matemática

Sejam:

x : o identificador da empresa para a qual deseja-se conhecer o grupo econômico;

$S = \{x_1, x_2, \ldots, x_n\}$: o conjunto de identificadores de todas as empresas;

$g(x) = \{g_1, g_2, \ldots, g_z | g_j \in S\}$: a função $g(x)$ que retorna o conjunto de identificadores do grupo econômico ideal para o identificador x. Isto é, o padrão-ouro esperado como resultado para uma consulta de x;

$f(x) = \{y_1, y_2, \ldots, y_m | y_i \in S\}$: a função $f(x)$ que seleciona automaticamente um subconjunto de identificadores de empresas em S que compõem o grupo econômico de x;

Deseja-se construir $f(x)$ de modo que seja o mais próximo possível de $g(x)$. Todas as métricas a serem extraídas serão comparando $f(x)$ à $g(x)$. Neste trabalho, mediu-se a eficiência de $f(x)$ em relação à $g(x)$ por meio das medidas de **precisão** e **revocação** que são detalhadas na seção 4.2.5 onde uma discussão sobre as métricas é apresentada em detalhes.

3. METODOLOGIA

A seguir é apresentada a metodologia que foi seguida para a execução desse trabalho.

3.1 Construção manual de grupos econômicos

Supondo que alguém deseje conhecer o grupo econômico de uma determi- nada empresa, uma das possíveis formas de construí-lo é por meio da base pública da Receita Federal. Nesta abordagem, o primeiro passo começa ao consultar o Cartão CNPJ[2] da empresa de interesse para, em seguida, acessar o seu cartão QSA (Quadro de Sócios e Administradores) que contém a lista de empresas que possuem sociedade direta com a empresa sendo pesquisada. Em caso de a empresa não possuir relação societária com outras empresas, o processo é encerrado. Caso contrário, para cada empresa da lista de sócias repete-se o processo anterior, recursivamente, construindo toda a estrutura societária até alcançar, em determinado nível, um conjunto de empresas que já não possuem sociedade com outras empresas. Até este ponto, tem-se um grupo de empresas que estabelecem relações societárias entre si, compartilham interesses mútuos e não competem diretamente. Em outras palavras, as empresas que possuem relações societárias entre si fazem parte de um mesmo grupo econômico.

Para exemplificar o processo descrito, realizou-se uma busca pelo cartão CNPJ da empresa *Mammoet Brasil Guindastes LTDA* de **cnpj: 11.426.377/0001–23**. A Figura 1 apresenta o cartão QSA referente à empresa *Mammoet Brasil Guindastes LTDA*, e o seu quadro societário composto por 2 sócios pessoa física e 2 sócios pessoa jurídica. O próximo passo será adicionar a estrutura de grupo econômico os 2 sócios pessoa jurídica encontrados na busca. Por fim, deve-se repetir o processo de consulta do cartão QSA para esses sócios pessoas jurídicas encontrados, adicionando

[2] https://servicos.receita.fazenda.gov.br/Servicos/cnpjreva/cnpjreva_solicitacao.asp

à estrutura de grupo econômico os novos sócios pessoas jurídicas (se houver) e repetir esse processo, recursivamente, até que todas as empresas alcançaveis a partir de *Mammoet Brasil Guindastes LTDA* tenham sido incorporadas a estrutura de grupo econômico sendo construída.

É importante enfatizar que neste trabalho não foi explorada nenhuma relação societária com pessoas físicas. Primeiro, porque o foco inicial eram as relações societárias apenas entre pessoas jurídicas, de modo a contrastar as informações presentes na base de dados da Receita Federal e as informações de grupos econômicos informados pelas empresas requerentes via formulário de notificação. Segundo, porque o trabalho específico com relações societárias envolvendo pessoas físicas é realizado, em grande medida, pelo projeto Cérebro (PIMENTA, 2019) desenvolvido pela Superintendência Geral do Cade.

Figura 1 – A empresa Mammoet Brasil Guindastes LTDA no cartão QSA

Consulta Quadro de Sócios e Administradores - QSA

CNPJ:	11.426.377/0001-23
NOME EMPRESARIAL:	MAMMOET BRASIL GUINDASTES LTDA.
CAPITAL SOCIAL:	R$20.175.446,00 (Vinte milhões, cento e setenta e cinco mil e quatrocentos e quarenta e seis reais)

O Quadro de Sócios e Administradores(QSA) constante da base de dados do Cadastro Nacional da Pessoa Jurídica (CNPJ) é o seguinte:

Nome/Nome Empresarial:	MAMMOET AMERICA SOUTH HOLDING B.V.		
Qualificação:	37-Sócio Pessoa Jurídica Domiciliado no Exterior	**País de Origem:**	PAÍSES BAIXOS (HOLANDA)
Nome do Repres. Legal:	CHRISTIAAN LAVOOIJ	**Qualif. Rep. Legal:**	17-Procurador
Nome/Nome Empresarial:	MAMMOET FINANCE B.V.		
Qualificação:	37-Sócio Pessoa Jurídica Domiciliado no Exterior	**País de Origem:**	PAÍSES BAIXOS (HOLANDA)
Nome do Repres. Legal:	CHRISTIAAN LAVOOIJ	**Qualif. Rep. Legal:**	17-Procurador
Nome/Nome Empresarial:	RICARDO FIORATTI COFFONE		
Qualificação:	05-Administrador		
Nome/Nome Empresarial:	ANDRES GOENAGA TRUJILLO		
Qualificação:	05-Administrador		

Fonte: https://servicos.receita.fazenda.gov.br/Servicos/cnpjreva/Cnpjreva_qsa.asp, verificado em 25 de julho de 2022

Da construção manual de grupos econômicos utilizando a base pública de dados da Receita Federal emergem três desafios principais: direcionalidade das relações societárias, complexidade de verificação e incomensurabilidade do esforço que serão descritas a seguir.

3.1.1 Direcionalidade das relações societárias

Iniciar a pesquisa por *Mammoet Brasil Guindastes LTDA* permite encontrar uma relação societária entre ela e *Mammoet America South Holding B.V*. porém, o inverso não é verdadeiro. Isto é, ao realizar uma consulta por *Mammoet America South Holding B.V.*, a empresa *Mammoet Brasil Guindastes LTDA* não é retornada como sócia, conforme pode ser verificado na Figura 2 que exibe o cartão QSA para a empresa *Mammoet America South Holding B.V.* (sem sócios). Isso acontece porque as relações societárias entre as empresas na base pública da Receita Federal são direcionadas. Ou seja, a consulta via cartão QSA apresenta unicamente uma lista de empresas que são sócias da empresa sendo pesquisada, não informando, contudo, as empresas das quais ela é sócia (*tem sociedade com*), por exemplo. Dessa forma, a orientação das relações societária impede que seja possível traçar um caminho inverso (*é sócia de*) e com isso conceber o grupo econômico a partir de qualquer empresa sendo pesquisada. Como consequência dessa constatação, é possível afirmar que o grupo econômico irá mudar a depender da primeira empresa sendo pesquisada, e isso compromete sobremaneira o processo de construção manual de grupos econômicos.

Figura 2 – Cartão QSA da Mammoet America South Holding B.V.

Fonte: https://servicos.receita.fazenda.gov.br/Servicos/cnpjreva/Cnpjreva_qsa.asp, verificado em 25 de julho de 2022

3.1.2 Complexidade de verificação

De que modo seria possível verificar se o grupo econômico construído manualmente está correto? Conforme argumentado anteriormente, devido a orientação das relações da base da Receita Federal, a construção do grupo é influenciada pela empresa semente, isto é, a primeira empresa a iniciar o ciclo de pesquisas. Mas ao final do processo, como verificar se o resultado final está correto ou minimamente coerente? Não há. Para o contexto específico do Cade, seria possível comparar o resultado do grupo econômico produzido ao grupo econômico apresentado pelas empresas via formulário de notificação. Nesse ponto, deve-se que admitir uma premissa

questionável a de que **as empresas sempre informam seus grupos econômicos corretamente**. Sendo a análise de grupos econômicos diretamente relacionada ao mercado relevante e a aplicação de multas e sanções, é natural que em vários cenários haja um incentivo por parte de algumas empresas a tentarem nublar as informações refentes aos relacionamentos societários de forma integral. Uma proposta alternativa para a construção de grupos econômicos de forma manual seria iniciar a busca por relações societárias para cada uma das empresas informadas via formulário de notificação, agrupando, ao final da pesquisa, todas as empresas encontradas em um mesmo grupo. Apesar de tentadora, essa abordagem descaracteriza o problema original que busca conhecer o grupo econômico recebendo como entrada um único identificador da empresa, e, neste caso, recebe-se como entrada não apenas o identificador de uma empresa, mas, uma lista de empresas informadas via formulário de notificação. Enquanto o problema original formulado na seção 2.3 extrapola sua utilidade para além das atividades do Cade, a análise de grupos econômicos dependente das informações do formulário de notificação como parâmetros de entrada minimiza esta utilidade que passa a ser restrita às atividades do Cade. Dessa forma, essa opção de reconsiderar o problema original não foi levada adiante.

3.1.3 Incomensurabilidade do esforço

Ao iniciar o processo de construção manual de grupos econômicos não é possível dimensionar o esforço necessário para a conclusão do trabalho. *Por quantos níveis deve-se repetir o processo? Quantas empresas serão listadas em cada nível?* Não é possível mensurar. O que gera um novo desafio: incomensurabilidade do esforço. Isto é, não se sabe, a priori, o tempo necessário para a conclusão da atividade, nem mesmo se ela é possível de ser concluída em tempo hábil. São $\sim$ 45.000.000 de empresas, $\sim$ 27.000.000 de relações societárias, qualquer empresa que possua uma fração, ainda que pequena desse volume já tornará o processo manual impraticável em tempo hábil.

3.2 Proposta de Solução

A seguir são apresentados os tópicos que envolveram a construção da solução.

3.2.1 Base de dados pública da Receita Federal

Para este trabalho foi utilizada a base de dados pública da Receita Federal[3] disponível. Aqui faz-se necessário distinguir o processo de construção de bases de dados entre as bases pública da receita em 2020 e 2021. A base pública da Receita Federal é geralmente distribuída por meio de

[3] https://www.gov.br/receitafederal/pt-br

20 arquivos compactados, cada um deles contendo, aproximadamente 5GB de dados. Entretanto, enquanto a base de dados disponibilizada em 2020 é distribuída por meio de um conjunto de linhas de 1200 caracteres onde cada variável precisa ser extraída por meio de um pré-processamento. A base de 2021 apresenta um *design* de estrutura de dados completamente diferente da de 2020. Não sendo possível, simplesmente aproveitar a primeira estrutura e povoá-la com dados de 2021, mas sim, refazê-la de forma independente. A Figura 3 apresenta um exemplo do conteúdo de uma linha do arquivo compactado em seu formato bruto. A forma de interpretá-la é descrita em detalhes no documento de layout fornecido pela própria Receita Federal.

Figura 3 – Exemplo de linha em estado bruto retirado dos arquivos da Receita Federal.

Fonte: Elaboração própria com dados da Receita Federal.

Inicialmente foi necessário construir um *schema* de banco de dados relacional responsável por armazenar todas as tabelas criadas seguindo a proposta de *layout* descrita anteriormente.

O banco de dados escolhido para este processo foi o *Postgres*. De forma geral, este banco de dados relacional é o responsável por manter os dados sob 4 propriedades fundamentais: **Atomicidade**, **Consistência, Isolamento** e **Durabilidade** (ACID) (JATANA et al., 2012).

- **Atomicidade:** permite tratar operações de forma transacionada. Isto é, ao executar alguma operação sobre os dados há apenas duas opções possíveis: ou tudo o que foi proposto é executado, ou nada é executado. Essa propriedade exclui a possibilidade de execução parcial, tornando o banco de dados capaz de recuperar-se ao seu estado original (*rollback*) sempre que algo der errado no meio do caminho. A título de exemplo, em caso de haver necessidade de alterar o endereço das empresas, removendo, por exemplo, todos os caracteres especiais de seus nomes, se em determinado ponto do processamento o fornecimento de energia é interrompido, não é possível saber, a priori, a quantidade de linhas que foram alteradas e o que falta para ser concluído. É

justamente esse impasse que pode comprometer a consistência das informações. O princípio da atomicidade garantirá que todas as alterações sejam realizadas ou que nenhuma alteração seja realizada retornando então a banco de dados para a sua versão original.

- **Consistência:** Uma relação é um conceito matemático no qual todos os elementos compartilham as mesmas propriedades. Uma tabela com duas colunas: nome e idade, por exemplo. Essa tabela constitui uma relação matemática na qual todos os elementos (linhas) possuem um nome e uma idade. Não deve ser possível inserir elementos que contenham, por exemplo, imagem quando deveria ser nome, ou texto quando deveria ser idade. A manutenção dessa propriedade assegura que todos os elementos de uma coluna, são de um mesmo tipo de dados o que gera confiança para extrair métricas sobre elas e realizar análises. Que confiança haveria nos resultados de um banco de dados após constatar-se que algumas idades possuem, a letra "a" ou números negativos "-78" ? Como ficariam as métricas? Os índices de tendência central? Os desvios-padrão? Muito provavelmente as análises estariam comprometidas.

- **Isolamento:** Essa propriedade garante que múltiplas operações possam ser realizadas no banco de dados sem que haja interferência entre elas. Isto é, enquanto consultas estão sendo executadas, alterações podem ocorrer paralelamente sem que o acesso ou a consistência dos dados sejam comprometidos. Esta propriedade é fundamental para permitir que o banco de dados permita múltiplos aces- sos aos dados mesmo tendo que realizar alterações nos dados ou na estrutura sem que para isso tenha que tornar o serviço inoperante para realizar esses ajustes.

- **Durabilidade:** As alterações propostas aos dados devem persistir mesmo em casos de quedas de energia, travamentos, falhas ou erros de quaisquer natureza. Após o sucesso da execução de uma operação o resultado da transação sobre os dados deve manter-se de forma permanente em memória não-volátil e em definitivo.

Muitas tecnologias recentes de bancos de dados, em especial as *NoSQL*, apesar de apresentarem melhor desempenho em cenários específicos, não garantem as propriedades apresentadas, sendo esta, a principal razão de não terem sido exploradas. Além dessas, há uma série de detalhes técnicos sobre bancos que dados relacionados ao trabalho desenvolvido que, neste momento, extrapolam o foco principal sendo apresentado. Autores tais como: (FONTAINE, 2020) e (EISENTRAUT; HELMLE, 2013) detalham tecnicamente os conceitos apresentados e apresentam-se como recursos para consulta dos conceitos descritos.

3.2.2 Comparativo descritivo entre junho de 2020 e maio de 2021 para a base da Receita Federal

Entre junho de 2020 e maio de 2021 ocorreram uma série de mudanças na dinâmica econômica do país. Muitas dessas mudanças foram refletidas na base de dados da Receita Federal. A seguir, são apresentadas comparações entre essas duas bases de informação:

- Entre junho de 2020 e maio de 2021 foi registrado um aumento de 111% na quantidade de empresas inaptas, 109% na quantidade de empresas baixadas e uma diminuição de 16% na quantidade de empresas ativas. Surgiram 4,5 milhões de novas empresas e 112 mil novas filiais. Isso é equivalente a um aumento de 11% a mais no número de empresas e 5% a mais no número de filiais;

- Apesar do aumento do número de empresas, entre julho de 2020 e maio de 2021 houve diminuição no número de relações societárias entre empresas. Antes havia pouco mais de 500 mil relações societárias e, agora, temos 272 mil (apenas entre empresas). Isso é o equivalente a dissolução de 45% das relações societárias entre empresas;

- Em junho de 2020 existiam 11 milhões de empresas com relações societárias com pessoas físicas. Em maio de 2021 passou-se a ter 10 milhões de empresas com essa configuração. Isso significa 10% menos relações societárias entre empresas e pessoas físicas. Também verificou-se que 97% das relações societárias são com pessoas físicas. Isto é, o relacionamento societário entre empresas é secundário;

- Varejo e Salão de Beleza são as atividades com mais empresas ativas. Ao serem comparadas, a base da Receita Federal entre junho de 2020 e maio de 2021 é possível perceber uma diminuição de empresas em todas as atividades com exceção de *Promoção de vendas e Condomínios Prediais*. Em junho de 2020 o top 10 das atividades com mais empresas ativas incluíam *Atividades de Associações de Defesa de Direitos Sociais* que foi superada em 2021 pela atividade de *Fornecimento de Alimentos Preparados Preponderantemente para Consumo Domiciliar*. A Figura 4 apresenta o histograma das top 10 atividades econômicas com mais empresas ativas no Brasil em 2021;

- Em junho de 2020 a empresa que mais possuía sócios pessoa-física era QUINTON PRONTO SOCORRO LTDA com 658 sócios. Enquanto a empresa que mais possuía sócios pessoa-jurídica era a UNIMED PARTICIPAÇÕES LTDA com 228 sócios. Em maio de 2021 a

empresa que possuía mais sócios pessoa-física era a OGS SERVICO DE ATENDIMENTO MEDICO-HOSPITALAR DE SÃO JOSé DOS CAMPOS LTDA com 952 sócios. Enquanto a empresa com mais sócios pessoa-jurídica era a CONSÓRCIO ALSOLAR com 359 sócios;

Figura 4 – Top 10 atividades com mais empresas ativas no Brasil em 2021

Fonte: Elaboração própria com dados da Receita Federal.

- Em junho de 2020 havia pessoas físicas com mais de 13.000 relações societárias. Já em maio de 2021 essas relações foram reestruturadas para menos de 2.000;
- Em junho de 2020 havia mais de uma centena de empresas com mais de 1.000 relações societárias. Em maio de 2021 foi possível notar uma reestruturação dessas relações societárias de modo que a empresa com mais relações societárias não ultrapassou 360 relações;
- Em maio de 2021 havia 15.564 empresas do tipo consórcio. São 6% a mais que em 2020.

3.2.3 Principais Desafios

A construção da base pública da Receita Federal apresentou inúmeros desafios, dentre eles é possível destacar:

1. **Definição imprecisa da codificação dos arquivos:** foram testados todos os tipos disponíveis e o que houve maior compatibilidade foi o tipo *latin1*. Ainda assim, vários ajustes precisaram ser feitos para que os dados fossem limpos e estivessem prontos para serem consultados;

2. **Dados inconsistentes com o layout:** muitos arquivos continham uma quantidade robusta de linhas impuras. Valores nulos, caracteres especiais, e vários elementos que feriam a consistência do banco e precisaram ser tratados antes de serem efetivamente incorporados ao banco de dados. A Figura 3, apresenta um exemplo de impureza próximo a palavra "Brasil", um caractere nulo, que impede que esta linha seja inserida no banco de dados sem a devida correção;

3. **Tamanho dos arquivos:** arquivos com 5 milhões de linhas não são manipulados de forma convencional porque a maioria dos editores de texto não são hábeis em fornecer edição para um arquivo dessa magnitude. Então, a simples tarefa de encontrar uma ferramenta com essas competências torna-se um desafio. Neste trabalho, sempre que houve a necessidade de analisar o interior dos arquivos, utilizamos o Sublime[4]. Se arquivos apresentem apenas 1% de impureza nos dados de modo que estas inconformidades precisem ser reparadas (ou descartadas) antes de serem armazenadas. Isso gera ~ 50.000 pontos de falhas. A solução manual torna-se impraticável em tempo hábil. Uma segunda opção é buscar automatizar o processo. Isto é, construir um algoritmo que encontre todos esses pontos de falhas e corrija-os automaticamente de forma iterativa e incremental. Infelizmente, essa proposta não resulta eficiente. Iterar de 1 por 1 verificando se determinada condição é satisfeita apesar de mais eficiente que a primeira, proposta manual, ainda é insuficiente para resolver o problema em tempo hábil. Isto é, supondo que o processador gaste 0,01s para verificar se uma determinada linha contém (ou não) uma anomalia que precise ser corrigida, para um arquivo de ~ 5.000.000 de linhas (formato padrão para cada arquivo da base de dados de 2020), levaria ~ 14 horas por arquivo. Uma vez que a base de dados completa é composta por 20 arquivos, seriam necessárias ~ 280 horas dedicadas apenas à correção de impureza dos arquivos.

   Uma das técnicas exploradas para a solução de problemas dessa natureza foi a construção de expressões regulares que se propõem a corrigir problemas em cadeias de caracteres e a sua execução direta e paralela via sistema operacional. Para isso foi utilizado o comando **sed** executado diretamente no sistema operacional unix da seguinte maneira:

```
1    $ sed -i 's/\\/\//g' *L.00004
```

Código 3.1 – Comando em bash utilizado para limpar arquivo de forma escalável.

[4] https://www.sublimetext.com/

O comando presente no Código 3.1, substitui todas as "\"por"/" em todo o arquivo de final L00004 de uma única vez sem a necessidade de iteração linha a linha.

4. **Transmissão de dados via VPN:** para permitir que a base da Receita Federal sendo construída pudesse ser acessada por múltiplos usuários, foi necessário a criação e configuração de uma máquina virtual dentro da infraestrutura do Cade. Após receber todas as permissões, todos os *scripts* e códigos foram enviados via VPN junto com todos os arquivos já corrigidos. O desafio ocorria sempre que alguma oscilação na rede VPN interrompia o envio e o arquivo tinha de ser reenviado. O tempo médio de envio de cada arquivo variava entre 10 e 12 horas. Para manter o fluxo contínuo de envios via VPN e não perder os envios sempre que a VPN fosse desconectada foi utilizado o software *WinSCP*[5] que permitia a recuperação de falhas e a continuação do envio dos arquivos após instabilidades na rede interna do Cade.

3.2.4 Construção de um algoritmo escalável

A ideia do algoritmo desenvolvido se assemelha ao processo de construção manual de grupos econômicos descrito na seção 3.1, porém, com uma etapa adicional. Dado um identificador de empresa, neste caso, o **cnpj**, realiza-se uma consulta, semelhante a realizada anterior ao cartão **QSA**, porém, neste caso, em duas direções. Deseja-se conhecer todas as empresas que estão relacionadas com a empresa alvo sendo pesquisada, tanto as que estão incluídas na sua composição societária, quanto aquelas em cuja composição societária ela está incluída. A Figura 5 apresenta um extrato da tabela de sócios da base pública da Receita Federal. A coluna cnpj, apresenta os **cnpjs** das empresas que possuem alguma relação societária. A coluna **cpf_cnpj**, apresenta o número de **cpf** ou **cnpj** de sócios da empresa. Sendo assim, ao consultar o cnpj de uma determinada empresa na coluna **cnpj**, é possível alcançar todas as pessoas físicas e jurídicas que são suas sócias. Adicionalmente, ao consultar o **cnpj** de determinada empresa na coluna **cpf_cnpj**, sabe-se as empresas das quais ela faz parte da estrutura societária.

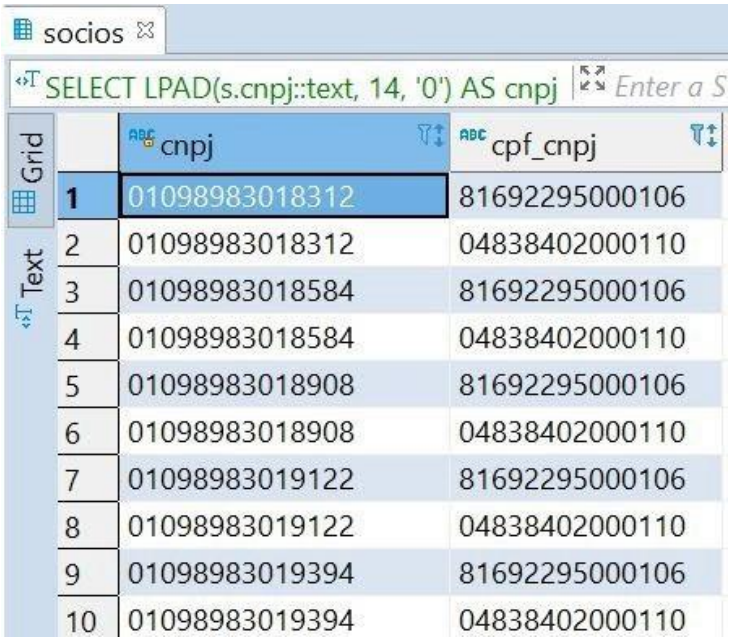
O algoritmo de construção automática de grupos econômicos baseia-se em agrupar todas as empresas alcançáveis pela coluna **cpf_cnpj** via coluna **cnpj**, recursivamente, em um único conjunto de empresas sócias, até que todas as empresas alcançáveis a partir de um **cnpj** inicial tenham sido processadas.

Dito de forma simples, busca-se por todas as empresas que são alcançáveis a partir de uma empresa alvo. Para isso, deve-se ignorar a direcionalidade das relações conforme apresentado na

[5] https://winscp.net

seção 3.1 tópico 3.1.1, e deve-se explorar todos os caminhos possíveis entre a empresa alvo e as demais empresas adjacentes. Isto é, todas as empresas para quais exista um caminho partido da empresa alvo, ou tendo-a como rota, farão parte do mesmo grupo econômico. A Figura 6 ilustra o processo de verificação de relações societárias entre empresas e a construção gradativa de grupos econômicos seguindo o algoritmo proposto.

Figura 5 – Porção de dados extraída da tabela de sócios da Receita Federal. A coluna cnpj contém o cnpj das empresas alvo, enquanto que a coluna cpf_cnpj contém o cnpj das empresas que são sócias das empresas alvo.


socios

SELECT LPAD(s.cnpj::text, 14, '0') AS cnpj

Grid | Text

	cnpj	cpf_cnpj
1	01098983018312	81692295000106
2	01098983018312	04838402000110
3	01098983018584	81692295000106
4	01098983018584	04838402000110
5	01098983018908	81692295000106
6	01098983018908	04838402000110
7	01098983019122	81692295000106
8	01098983019122	04838402000110
9	01098983019394	81692295000106
10	01098983019394	04838402000110

Fonte: Elaboração própria com dados da Receita Federal.

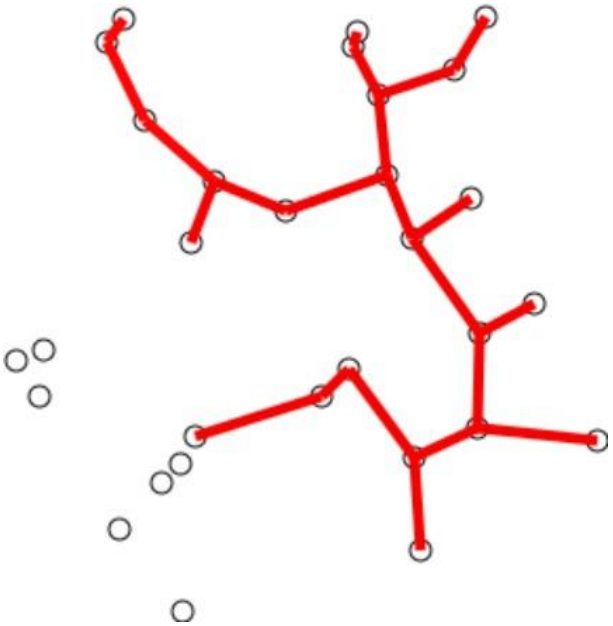
Até o momento discutiu-se com maiores detalhes **o que** foi feito desde uma perspectiva teórica, que em linhas gerais foi substituir a orientação da relação **A → B**, onde **A** é sócio de **B** de forma unidirecional, por uma relação biunívoca **A ←→ B**, onde A possui uma relação societária com **B** e **B** também possui uma relação societária com **A**. Esse procedimento é repetido até que todas as empresas alcançáveis a partir de uma empresa alvo sejam alcançadas, então, define-se o resultado obtido como sendo o grupo econômico.

A partir deste ponto será discutido o como foi realizado a construção do algoritmo de grupos econômicos, desde uma perspectiva técnica, com maiores detalhes de implementação. Em linhas gerais, há três caminhos para implementar o algoritmo proposto:

Grafos: A teria dos grafos, trata o problema em questão desde uma perspectiva de rede. A tabela de sócios da Receita Federal é vista como grafo, cada **cnpj** é visto como um vértice e

cada relação societária entre as colunas **cnpj e cpf_cnpj** é vista como uma aresta. A Figura 7 apresenta um exemplo da representação gráfica da estrutura matemática de um grafo. Nesta perspectiva, tanto o algoritmo de busca em largura (BFS), quanto o busca em profundidade (DFS) ou quaisquer algoritmos relativos à conectividade em grafos, tais como Prim em (BORUˇVKA, 1926) e (PRIM, 1957), Dijkstra (DIJKSTRA et al., 1959) ou Bellman-Ford (BELLMAN, 1958) seriam suficientes para serem minimamente ajustados de modo a produzirem como saída o resultado esperado de um grupo econômico.

Figura 6 – Exemplo de execução de algoritmo de Prim.

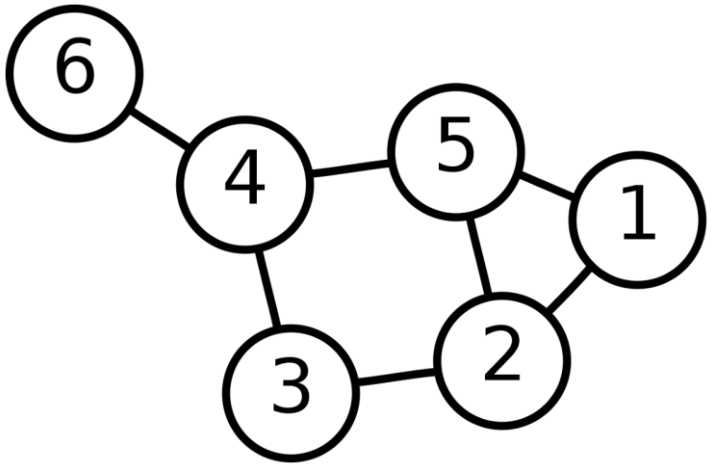


Fonte: (JI, 2022)

Álgebra Relacional: A álgebra relacional trata o problema desde uma perspectiva da teoria dos conjuntos. A Figura 8 apresenta 4 conjuntos relacionados em transitividade por meio de suas intersecções. Ou seja, para o caso em que o conjunto **A** contenha as empresas que possuam alguma relação com uma empresa **x** (sendo sócia de **x**, ou tendo **x** em sua estrutura societária). Por sua vez, que o conjunto B contenha empresas que possuam alguma relação com uma empresa **y**. Então, se **A** ∩ **B**≠ **Ø**, significa que há ao menos uma empresa que pertence à **A** e **B** simultaneamente, e, nesse caso, é possível agrupar as empresas dos conjuntos **A** e **B** uma vez que fazem parte do mesmo grupo econômico. O processo é repetido, recursivamente, até que todas as intersecções existentes sejam esgotadas. Da mesma maneira que o mesmo problema matemático pode ser resolvido analítica ou geometricamente, o que está sendo feito, neste caso, é valer-se da álgebra relacional para buscar empresas alcançáveis a partir da empresa alvo. É uma nova

abordagem teórica capaz de produzir o mesmo resultado esperado.

Figura 7 – Exemplo de grafo onde os círculos são os vértices, cada número representa um identificador da entidade e as retas são as arestas, isto é, as relações entre os vértices.



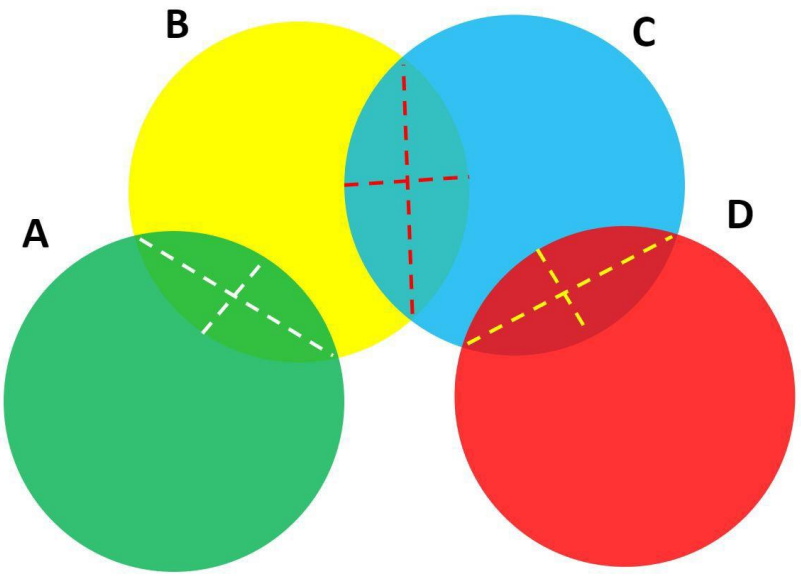
Fonte: (TOTH, 2022)

Clustering: A solução de *clusters*, assemelha-se a proposta de teoria dos grafos, porém com duas diferenças principais: primeiro, porque o algoritmo irá explorar apenas um número limitado de níveis, ao invés de avançar até encontrar em- presas sem sócios ou que não compõem sociedade. Segundo, porque é possível definir a priori o número de grupos que se deseja obter ao final do processo. A Figura 9 apresenta um exemplo de *cluster* definido sob um grafo direcionado. Nesta perspectiva, ambos os parâmetros (quantidade de grupos, número de níveis) são passíveis de configuração. É importante enfatizar que essa proposta não chegou a ser implementada, sendo melhor descrita na seção 6.1 onde são apresentados alguns dos possíveis trabalhos futuros.

A primeira solução que foi implementada seguia o padrão da álgebra relacional por meio de operações recursivas. O Código 3.2 apresenta um exemplo de possível implementação do algoritmo proposto na ferramenta *Postgres* para o **cnpj** (identificador da empresa) de número **012.59.348.0001-60**. Este enfoque apresenta vantagens. Inicialmente, a implementação do algoritmo ocorre na mesma linguagem (SQL), a mesma linguagem em que os dados foram estruturados e consultados. Isso significa que não há necessidade de investir tempo na aprendizagem de novas tecnologias, contratação de mão de obra especializada, tampouco de inserir novas ferramentas dentro da arqui- tetura de solução proposta. Por outro lado, na prática, essa solução resultou ineficiente. A construção de cada grupo econômico levava em média 30s para ser completada e isso tornava a proposta de solução impraticável de ser executada em um

contexto real. Um outro problema crônico de soluções recursivas é a sobrecarga de utilização de memória. A depender da quantidade de relações que uma empresa tivesse ou a quantidade de níveis necessários à percorrer até a conclusão do algoritmo, o consumo de memória aproximava-se do colapso. Isto é, uma falha conhecida como *stackoverflow*, ou estouro de pilha (FRONTIER, 2022). É necessário relembrar que existem ~ 29.000.000 de relações societárias que precisariam ser analisadas e tanto eficiência em termos de tempo quanto a escala são fundamentais para o êxito da solução.

Figura 8 – Exemplo de verificação de intersecção implementada de forma recursiva via álgebra relacional.

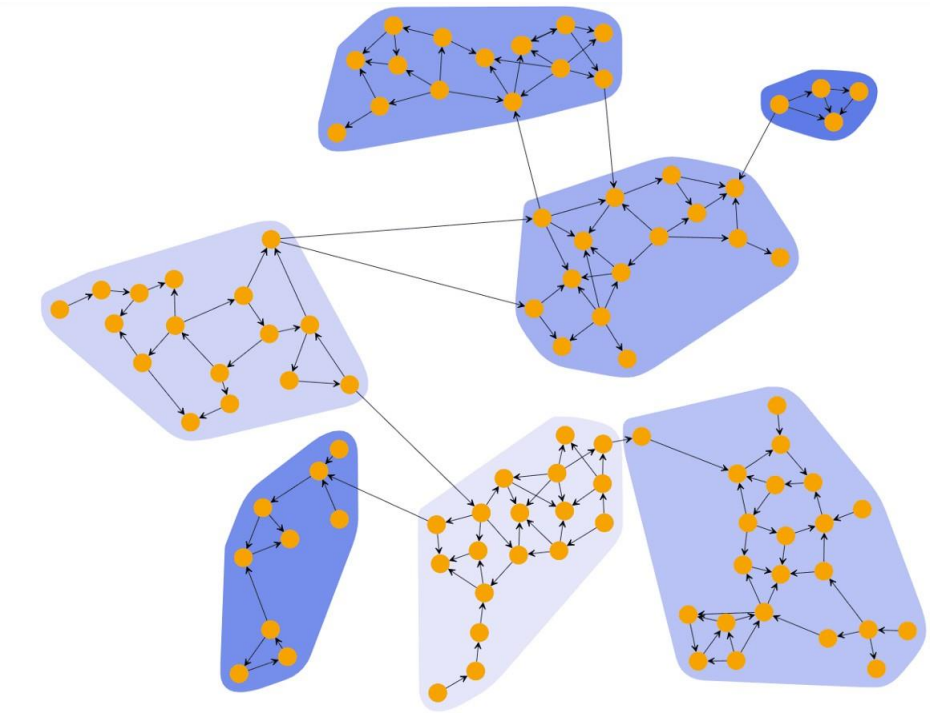


Fonte: Elaboração própria.

```
WITH RECURSIVE grupo_economico(cnpj, nome) AS (
    SELECT s1.cnpj, CAST(s1.nome_socio AS TEXT) AS nome
    FROM receita.socios s1
    WHERE s1.identificador_de_socio = 1 AND
          s1.cnpj = 01259348000160
UNION    SELECT    CAST(s2.cpf_cnpj    AS    bigint),
         (grupo_economico.nome || s2.nome_socio) AS nome
    FROM receita.socios s2
    INNER JOIN grupo_economico
    ON    (s2.identificador_de_socio    =    1    AND
      CAST(s2.cpf_cnpj AS bigint) = grupo_economico.cnpj)
    WHERE s2.identificador_de_socio = 1)
```

Código 3.2 – Construção recursiva utilizando álgebra relacional no Postgres.

Figura 9 – Exemplo de construção de clusters a partir de um grafo orientado.



Fonte: (YWORKS, 2022).

Um algoritmo escalável foi alcançado combinando duas ferramentas: *Postgres* e *Python* dentro de uma série de fatores. Dentro da ferramenta Postgres foi realizada a filtragem de dados, indexação de dados e configuração de parâmetros internos. A seguir é ampliada a discussão sobre cada um dos pontos levantados:

Filtragem de dados: filtrar dados é uma abordagem interessante por reduzir a quantidade de dados que será processada. Algumas vezes ela é suficiente para transformar um processo impraticável em não apenas praticável como eficiente. O raciocínio é simples, após a filtragem, a quantidade de dados é menor o suficiente para que a massa de dados resultante seja executada dentro da infraes-trutura disponível em um tempo aceitável para o cliente. Em linhas gerais, a filtragem de dados pode ocorrer tanto verticalmente quanto horizontalmente. Isto é, reduz colunas e reduzir linhas, respectivamente. A filtragem realizada nesta etapa contemplou as duas possibilidades. Primeiro selecionando apenas duas colunas (**cnpj** e **cpf_cnpj**). Segundo, selecionando apenas empresas **matrizes**, **ativas**, **não-consórcios**. Sobre empresas matrizes, foi admitido que as filiais seguirão as relações societárias definidas pelas matrizes e que não é possível uma relação societária existir exclusivamente com uma filial sem que isso não seja aprovado pela matriz. Neste caso, foi

inferido que o comportamento das matrizes será o mesmo que o das filiais por transitividade. Também foram desconsideradas todas as relações existentes entre empresas com situação cadastral diferente de ativa. Quaisquer outras situações, tais como: nula, baixada, inapta ou suspensa, por exemplo, foram removidas da análise. Esse passo é importante porque elimina inúmeras relações entre empresas provocando, dessa forma, o aumento da coalescência[6] que tecnicamente resulta em menos grupos econômicos, porém mais densos.

Por fim, também foram excluídas todas as empresas de natureza jurídica do tipo *joint venture*. As empresas do tipo joint venture são fruto da união entre duas ou mais empresas com o objetivo principal de viabilizar um negócio por um período de tempo determinado, visando, por óbvio, o lucro. Em grande medida, foi possivel observar empiricamente que joint venturies unem grandes conglomerados e geram grupos massivos, porém, descontextualizados. Vale e Petrobrás, por exemplo, se relacionam por meio de um consórcio. Por meio desse vínculo, cria-se um grupo econômico com $\sim$ 80.000 empresas. Muitas dessas empresas, apesar de fazerem parte desse grupo, estão hierarquicamente tão distantes de Vale e Petrobrás que em alguns casos põe a dúvida se realmente fazem parte do mesmo grupo econômico. Com a intenção da construção de grupos mais coerentes e melhor definidos, decidiu-se por excluir as empresas de natureza jurídica do tipo *joint ventures*. Obviamente essa discussão pode ser amplamente estabelecida e revisada quantas vezes se fizer necessário, e, o filtro removido ou ampliado a depender do interesse específico da gestão.

Indexação de dados: Todos os dados foram estruturados e indexados conforme sugerido em (GARCIA-MOLINA; SALEM, 1992) e (PLATTNER et al., 2013).

Configuração de parâmetros internos: Todos os parâmetros internos do *Postgres* foram ajustados por meio do ***pgtuner***[7] para obter a máxima performance das consultas dentro da infraestrutura fornecida pelo Cade. Esse passo é importante para que seja possível extrair o melhor desempenho da estrutura de hardware fornecida, e com isso, potencializar o resultado das consultas.

---

[6] Segundo o dicionário online priberam a palavra coalescência pode ser entendida como sendo o processo de aderência de partes que se achavam separadas. Aglutinação. No contexto que vem sendo debatido, a coalescência pode ser entendida como a capacidade de agrupar empresas em um grupo. Quanto maior a coalescência, menor a quantidade de grupos e maior a quantidade de empresas em cada grupo.

[7] https://pgtune.leopard.in.ua/

Em *Python*, foi utilizada a implementação do BFS presente no pacote *NetworkX* recomendado por (HAGBERG; SWART; CHULT, 2008) devido a eficiência em manipulação em estruturas de grafos. No *Python*, é realizada uma consulta, já indexada, de todos os pares (empresa, sócia) para todas as empresas matrizes, ativas e não-*joint venture*.

Obtem-se então, como resultado, algo semelhante ao apresentado na Figura 5. Em seguida, é executado o algoritmo BFS do pacote *NetworkX* para encontrar todos os caminhos societários entre empresas. Por exemplo, existindo as relações (*A, B*), (*B, F*), (*F, H*), o caminho A ↔ B ↔ F ↔ H é criado. Todas as entidades de um mesmo caminho recebem um número associado que identifica o grupo a que faz parte. A Figura 10 apresenta um extrato da tabela da construção de grupos econômicos proposta para este trabalho. A tabela final é composta por duas colunas **cnpj** e **grupo** onde o cnpj representa o identificador único da empresa e o grupo o identificador do grupo no qual ela foi agrupada. Empresas que possuem o mesmo identificador de grupo fazem parte do mesmo grupo econômico. Toda a estrutura foi indexada para que fosse possível, e com eficiência, encontrar todas as empresas de um determinado grupo, como também verificar se determinada empresa faz ou não parte de um determinado grupo.

Figura 10 – Extrato da tabela grupos econômicos gerada ao final do processo.

Grid | Text

	cnpj	grupo
1099	33,098,658,124,537	1
1100	33,098,658,185,161	1
1101	33,098,658,081,544	1
1102	33,098,658,022,963	1
1103	33,098,658,079,566	1
1104	33,098,658,124,707	1
1105	14,146,966,000,146	1
1106	33,098,658,059,298	1
1107	33,098,658,062,329	1
1108	33,098,658,020,758	1
1109	31,973,293,000,117	1
1110	20,993,615,000,173	1
1111	33,098,658,079,990	1
1112	1,392,000,103	2
1113	56,724,412,000,129	2
1114	32,284,684,000,197	3

Fonte: Elaboração própria.

Por fim, após serem implementadas as estratégias descritas anteriormente tanto no banco de dados por meio das configurações e filtragens, quanto com o processamento de parte dos

dados em memória, foi possível sair de um universo de ∼43.000.000 empresas para ∼ 455.496 que precisavam ser analisadas. A performance foi superada de 30s para geração de cada grupo para 66, 085s para geração de todos os grupos (0, 085s consulta no banco + 66s da construção dos grupos via BFS). Isso possibilitou não apenas a possibilidade de conclusão do algoritmo, mas a possibilidade de que fosse escalável.

4. VERIFICAÇÃO E VALIDAÇÃO

Verificar e validar são coisas distintas. Enquanto a verificação analisa se o que foi produzido atende as especificações do que foi projetado. Validar consiste em medir se o que foi construído é capaz de resolver o problema. Verificar este trabalho consistiu em contabilizar, se foi possível conceber (ou não) um algoritmo automático que pudesse, mediante um identificador de empresa, produzir como saída seu grupo econô- mico, em tempo aceitável e escalável. Validar, por sua vez, consistiu em medir se os grupos econômicos encontrados podem ser utilizados no transcurso laboral, traduzindo o esforço computacional em concebê-los, em ganho de alguma eficiência pelo Cade.

Nessa seção, serão discutidos aspectos da verificação e validação propostas para esse trabalho.

4.1 Verificação

Para a verificação desse trabalho foram construídos manualmente vários cenários com as informações retiradas da própria Receita Federal. Em seguida, os resultados produzido pelo algoritmo foram comparados com os construídos manualmente. Em todos os casos, o resultado esperado pelo algoritmo foi 100% compatível com os cenários construídos manualmente. A Figura 11 apresenta um exemplo da construção de cenário manual para análise de conformidade do algoritmo.

4.2 Validação

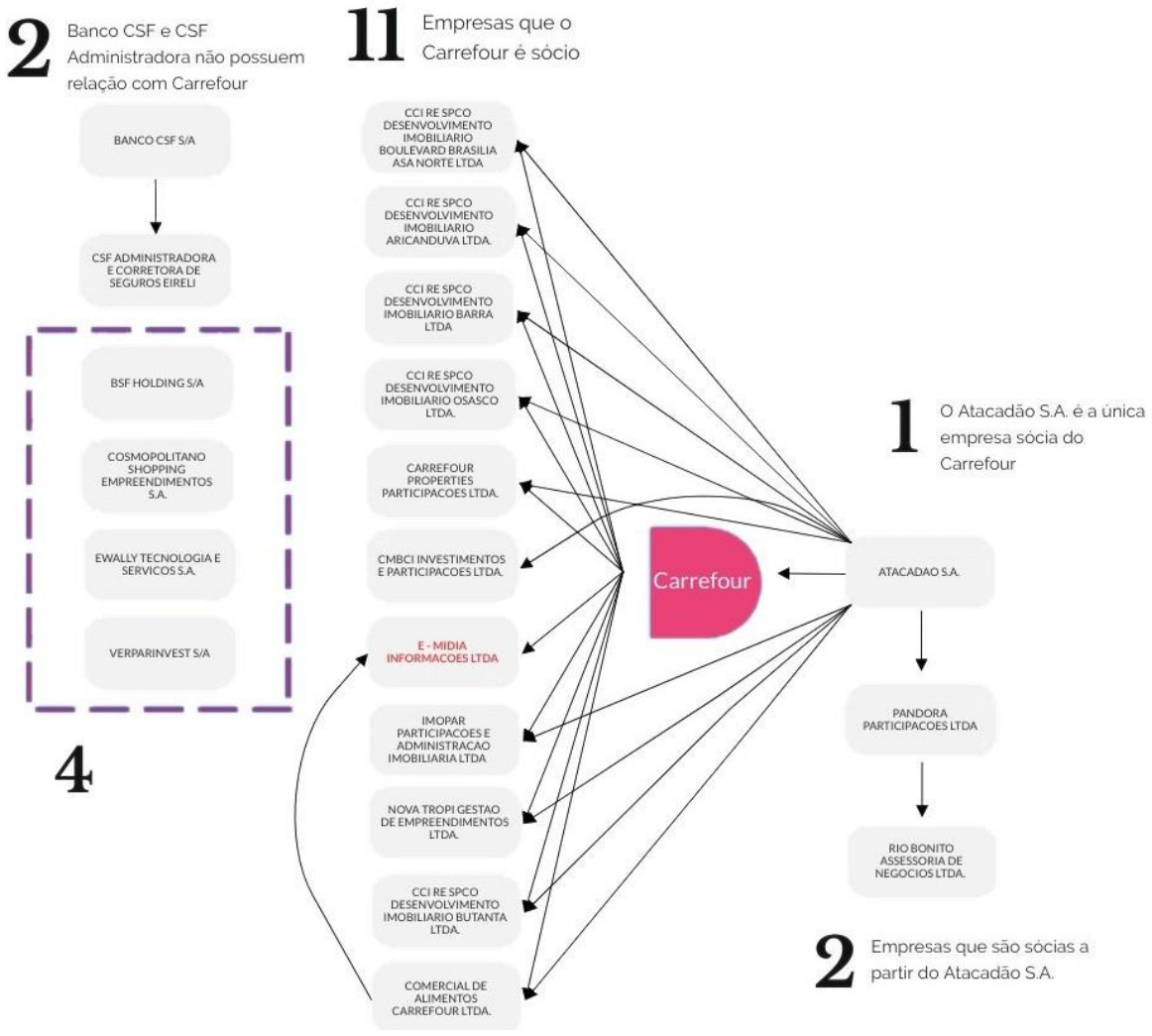
A seguir será descrito o processo de validação deste trabalho. Este processo inclui a construção do *benchmark* e exposição de suas dificuldades, a dificuldade envolvida com o trabalho manual e os principais problemas enfrentados.

4.2.1 Construção de um Benchmark

Como medir a qualidade da solução produzida? Quais suas limitações? Como medir sua utilidade? Para ter parâmetros capazes de auferir a qualidade do que foi produzido, bem como suas limitações, foi construído um *benchmark* a partir de todos os formulários de notificação

submetidos ao Cade pelas empresas interessadas em Atos de Concentração em dois períodos distintos: julho de 2020 e janeiro a março de 2021.

Figura 11 – Exemplo da construção de estrutura manual de relacionamentos para a empresa Carrefour.



Fonte: Elaboração própria. Informações retiradas da versão pública do AC de nº 08700.003654/2021-42.

Para cada processo, foram armazenados: o número do processo, as empresas interessadas no ato de concentração, e o nome das empresas que foram citadas como pertencentes ao grupo econômico de cada uma das interessadas, informações obtidas, em grande medida, do item II.5 do formulário de procedimentos ordinários submetidos ao Cade durante a submissão do AC. Ao todo, foram analisados manualmente 233 processos. Sendo estes 42 processos em 2020 e 191 processos em 2021. De cada processo foram extraídos ao menos 2 empresas. Foram excluídos todos os casos de acesso restrito e empresas estrangeiras sem cnpj presente na base de dados da Receita Federal analisada. Ao final, foram catalogados e analisados 286 grupos econômicos.

De modo a minimizar falhas humanas durante o processo de coleta de informações dos processos, esta atividade foi realizada de forma manual com validação cruzada em 4 etapas. A saber:

1. Todos os processos foram divididos igualmente entre 3 estagiários;
2. Cada integrante coletou os seguintes atributos dos processos: número do processo, as empresas interessadas, e as empresas que compõem cada um dos grupos econômicos relacionados às empresas interessadas presentes no item II.5 e/ou II.9 do formulário de notificação do processo;
3. Ao final, cada integrante recebeu uma nova tarefa, agora de conferência de uma porção do trabalho diferente da que havia trabalhado na etapa anterior para verificar se havia alguma inconsistência;
4. O item 3 foi repetido até a atividade alcançar uma condição de concordância unânime entre todos os envolvidos.

4.2.2 Análise descritiva dos problemas encontrados

Ao todo foram pesquisados 6.606 cnpjs de empresas presentes nos grupos econômicos de ACs selecionados entre 2020 e 2021. Especificamente, 842 empresas em 2020, no momento inicial da pesquisa e 5.764 em sua ampliação mais robusta realizada em 2021. De todos os cnpjs pesquisados, 427 empresas pesquisadas, o equivalente à 12% das empresas coletadas apresentaram algum problema durante o processo de coleta e verificação do cnpj. A seguir, são apresentados maiores detalhes sobre os problemas encontrados:

- 47% dos problemas encontrados (203 empresas) correspondem à empresas estrangeiras sem cnpj cadastrado na base da Receita Federal;
- 1% das empresas possuíam acesso restrito e não foi possível extrair maiores informações;
- 7% das empresas possuíam status diferente de ativo na base da receita, sendo listadas nos grupos econômicos estando: baixadas, inaptas ou nulas;
- 45% dos problemas corresponderam a nomes de empresas presentes nos formulários de notificação dos ACs mas que não foram encontrados na base da receita ou em sucessivas buscas pela internet. Esse problema poderia ser sanado caso o item II.5 do

formulário de notificação contemplasse, além da denominação, o cnpj de todas as empresas do grupo econômico.

4.2.3 Considerações sobre o trabalho manual

Um questionamento natural que poderia ser feito a esta altura é o seguinte: "*Por que não automatizar o processo de obtenção de grupos econômicos dentro dos formulários de procedimentos ordinários ao invés de valer-se de trabalho manual dos colaboradores?*" Em linhas gerais, a resposta a esta pergunta vem de dois argumentos principais: estrutura de arquivos e gênesis de verificação.

4.2.3.1 Estrutura do Arquivo

Todos os processos de AC submetidos ao Cade são construídos com plena liberdade de escrita. Cada escritório de advocacia preenche o formulário à sua maneira e essa ampla liberdade de estrutura e formatação faz com que os arquivos gerados sejam, naturalmente, não-estruturados. Faz-se necessário enfatizar que o termo não-estruturado, no contexto em questão, não significa, em absoluto, sinônimo de desorganizado, mas sim, fora de uma relação matemática bem definida, da qual todos os elementos possuem as mesmas propriedades, sendo fácil sua extração e processamento. Se fossem comparadas, a título de exemplo, duas estruturas: um texto livre, e uma tabela. A tabela, é um tipo estruturado. Todos os seus elementos possuem necessariamente as mesmas propriedades. Cada linha contém a informação completa sobre um único item enquanto que cada coluna, contém toda informação conhecida sobre uma variável ou atributo. Por outro lado, um texto corrido, é em geral, não-estruturado, não há um formado definido ou marcação especial que indique como extrair as informações. Como saber automaticamente dentro de um texto quem são os advogados? Quem são as empresas envolvidas? Quem são as empresas que fazem parte de um grupo econômico? Não havendo marcação especial ou estrutura que nos permita recuperá-los automaticamente faz-se necessário a construção de complexas expressões regulares e estruturas semânticas para empreender o feito. A complexidade em conceber uma solução para a extração automática de itens dentro dos processos (arquivos não-estruturados), extrapolaria o tempo necessário para a conclusão da tarefa, e desviaria, do objetivo principal desde trabalho, dessa forma, mesmo sendo possível de ser implementado no futuro, não foi realizado. Além disso, o Cade vem trabalhando em uma nova proposta de formulário eletrônico de procedimentos ordinários que provavelmente solucionará muitas questões discutidas, estruturando grande parte da informação necessária, inviabilizando, dessa forma esforços que migrem para o enfoque do não-estruturado.

4.2.3.2 Gênesis de Verificação

Supondo, em um cenário hipotético, que tivesse sido construído o algoritmo que automaticamente extraísse os grupos econômicos informados nos formulários de procedimentos ordinários preenchidos pelas empresas interessadas nos ACs e entregues ao Cade, e dessa forma, tivesse sido eliminada a necessidade de quaisquer interferência manual do processo de coleta dessas informações. De que forma seria possível verificar se o algoritmo em questão trouxe exatamente os grupos econômicos informados? Qual foi o seu percentual de erro? Para quais casos funcionou melhor e para quais casos foi insuficiente? A resposta a essas perguntas só seria possível confrontando as informações obtidas por este algoritmo contra um *benchmark* que contivesse um gabarito de valores esperados. A questão é que este *benchmark* inicial não existe e a sua construção inicial, constitui a **gênesis de verificação**. Este gabarito deve ser construído manualmente por especialistas de modo a concebê-lo de forma ideal. Ele constituirá o padrão-ouro, terá a premissa de corretude sobre a qual todas as demais métricas serão extraídas e discutidas, servindo, não apenas para a verificação do algoritmo hipotético, mas para todas as demais iniciativas nessa linha de estudos.

4.2.4 Principais Desafios Enfrentados

Durante o processo de construção do *benchmark*, muitos desafios que sequer foram previstos tiveram de ser superados. Essa seção, discute alguns desses desafios e alternativas que foram discutidas para superá-los.

4.2.4.1 Falsos cognatos

Ao pesquisar pela empresa “SEARA ALIMENTOS” na base da Receita Federal, muitas empresas com nomes similares são encontradas. Dessa forma, como identificar a empresa correta que está sendo citada no formulário de notificação? É importante enfatizar que ao referenciar a empresa errada, todo o grupo econômico relacionado estará comprometido. O formulário de notificação, por si só, não apresenta a obrigatoriedade de apresentar identificador único da empresa, no caso o **cnpj**. A Tabela 1 apresenta um exemplo de resultado ao buscar empresas pelo nome “SEARA”.

Tabela 1 – Exemplo de busca contendo inúmeros falsos cognatos para uma busca pela empresa “SEARA”

CNPJ	Nome
02.914.460.0112-76	SEARA ALIMENTOS LTDA
83.044.016.0030-68	SEARA COMERCIO DE ALIMENTOS LTDA

03.777.298.0001-39	PEREIRA SEARA COMERCIO DE ALIMENTOS LTDA
24.982.527.0001-27	SUPER SEARA COMERCIO DE ALIMENTOS LTDA

Fonte: Elaboração própria.

4.2.4.2 Erros Ortográficos

Erros presentes na escrita dos nomes das empresas dificultaram a sua identificação. A Figura 12 apresenta um exemplo em que as informações fornecidas pelas empresas contém erros ortográfico que dificultam ou impedem a sua identificação. Neste caso específico a empresa ***Thrown Nutrition*** não foi encontrada na base da receita, a empresa com maior proximidade léxica encontrada foi: ***Throuw Nutrition***.

Figura 12 – Exemplo de informação prestada pela empresa com erro ortográfico.

Segue abaixo lista das demais empresas pertencentes ao **Grupo SHV** com atividades no Brasil:

	Empresa
1.	Supergasbras Energia Ltda.
2.	Minasgas S.A. Indústria e Comércio
3.	Thrown Nutrition Brasil Nutrição Animal Ltda.
4.	Mammoet Brasil Guindastes Ltda.
5.	NNC Participações Ltda.
6.	SP Participações Ltda.
7.	SS Participações Ltda.
8.	Makro Food Service Ltda.

Fonte: Adaptado da versão pública do ato de concentração 08700.007277/2018-15, página 14.

4.2.4.3 Ininteligibilidade de nomes de empresas

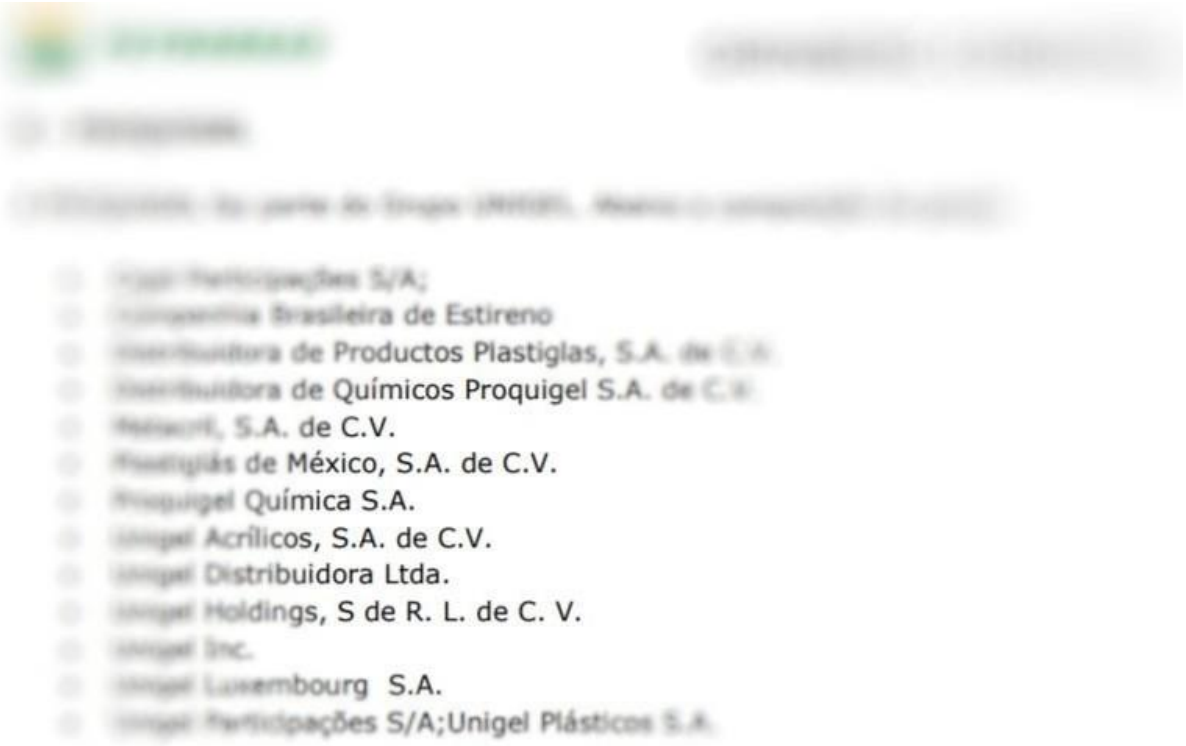
Algumas vezes, o excesso de abreviações impede a identificação da empresa e dessa forma, não é possível analisar corretamente o grupo econômico informado via formulário de notificação. A Figura 13 apresenta um exemplo de informações apresentadas contendo nomes de empresas ininteligíveis, como por exemplo: "**S de R. L. de C. V.**"[8].

Devido aos problemas expostos, foi necessário buscar manualmente pelo **cnpj** de todas as empresas presentes no *benchmark* para que a comparação se desse por **cnpj** e não mais por nome da empresa.

[8] A Figura 13 foi ofuscada propositalmente para não exibir a lista completa de empresas que fazem parte do grupo econômico.

Todo o processo de construção do *benchmark* utilizou a combinação de três técnicas: Computação por Humanos (QUINN; BEDERSON, 2011), *Mechanical Turk* (ROUSE, 2015) e Análise Delphi (HELMER, 1967). Pela *Mechanical Turk*, tem-se que todos os colaboradores envolvidos na construção foram estagiários do Cade com formação e experiência nas áreas de economia e direito, amplamente confortáveis com a manipulação dos processos e plenamente hábeis para realização das atividades, o que minimiza o ruído de interferência humana por inabilidade da atividade. Pela análise Delphi, é dado que todos os **cnpjs** inseridos foram verificados em duplicidade e em caso de divergência entre o **cnpj** e o nome da empresa houve uma terceira conferência e discussão pela unanimidade.

Figura 13 – Exemplo de Ininteligibilidade para o excesso de abreviações.



Fonte: Adaptado da versão pública do ato de concentração 08700002024/2020-70, página 9.

Para algumas empresas, o **cnpj** não foi encontrado, e nesses casos, de modo a não enviesar a contabilização de resultados extraídos com o *benchmark* esses valores foram removidos. A Figura 14 apresenta um exemplo de uma porção do *benchmark* construído. A primeira coluna contém o nome da empresa tal qual aparece no formulário de notificação do Cade, no item II.5. A segunda coluna, contém o cnpj para o nome da empresa que foi pesquisado em várias bases de dados na internet. A terceira coluna contém o cnpj-raíz que correspondem aos 8 primeiros dígitos do cnpj. A quarta coluna contém uma cor indicando a dupla verificação. Isto é, outro integrante,

verificou o nome da empresa e chegou ao mesmo cnpj, também verificou o status do cnpj e encontra-se como ativo, pertencente a uma empresa matriz do tipo não-consórcio. Por fim, a última coluna, contém uma observação que justifica alguma inconformidade com o cnpj encontrado e que precisa ser analisada com mais critério pelos integrantes do projeto.

Figura 14 – Exemplo de uma porção do benchmark desenvolvido.

08700000585/2020-34 MAFRA HOSPITALAR	12420164000157	12420164	
MAFRA HOSPITALAR SA	12420164000157	12420164	
HEALTH LOGISTICA SA (ATUAL DENOMINACAO DA CM LOGISTICA HOSPITALAR S	18320396000110	18320396	
TECNOCOLD LOCACAO DE ESPACOS E DISTRIBUICAO DE PRODUTOS REFRIGE	4212286000120	4212286	
CREMER SA	82641325000118	82641325	
CREMER ADMINISTRADORA DE BENS LTDA	12980235000176	12980235	
CM HOSPITALAR SA	12420164000157	12420164	
CCM INDUSTRIA E COMERCIO DE PRODUTOS DESCARTAVEIS SA	12288046000137	12288046	
CAMT EMPREENDIMENTOS E PARTICIPACOES LTDA	12225376000183	12225376	baixada
CIRURGICA MAFRA LTDA		1310222	
CMI HOSPITALAR LTDA	13809001000123	13809001	
CM CAMPINAS MEDICAMENTOS ESPECIAIS LTDA	11371888000195	11371888	
CM MEDICAMENTOS ESPECIAIS LTDA	4127483000140	4127483	
JMT EMPREENDIMENTOS E PARTICIPACOES EIRELI	29335401000184		
MAFRA LOCACOES LTDA	1310222000173	1310222	
MAFRA IMPORTACAO INDUSTRIA E COMERCIO DE PRODUTOS HIGIENICOS LTDA	9419262000160	9419262	
DCP PARTICIPACOES LTDA	30372013000152	30372013	
TRIO PARTICIPACOES LTDA	23493706000138	23493706	

Fonte: Elaboração própria.

4.2.4.4 Empresas estrangeiras em ACs nacionais

Inúmeras empresas estrangeiras não possuem cnpj no Brasil mas submetem atos de concentração em território nacional, dessa forma, não é possível associar, a partir das proponentes, o respectivo grupo econômico uma vez que não há referência para essas empresas na base pública da Receita Federal.

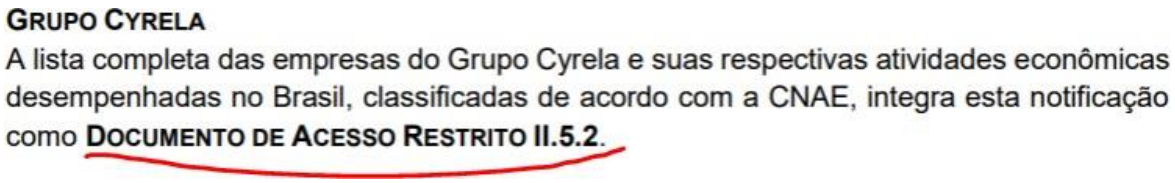
4.2.4.5 Dados de acesso restrito

Em inúmeros processos as informações referentes à grupos econômicos estão restritas ou incompletas, isso impacta diretamente nas métricas de precisão e revocação. Isto é, não é possível de saber se o algoritmo acertou ou errou sua construção de grupos econômicos simplesmente porque não há acesso ao grupo econômico completo. A Figura 15 apresenta um exemplo de grupo econômicos restrito. Neste caso, o grupo Cyrela solicitou sigilo às informações apresentadas para o seu grupo econômico. E, dessa, forma, não há como medir, sem acessar os documentos restritos, se o grupo econômico encontrado pelo algoritmo é compatível ou não com o apresentado pelas requerentes via formulário de notificação. Essas informações podem ser verificadas no processo de número 08700.001618/2021-44.

4.2.4.6 Inconsistência das informações apresentadas em diferentes processos

Há três tipos de inconsistências importantes que devem ser consideradas: a inconsistência cruzada, a inconsistência sequencial e a inconsistência de situação cadastral.

Figura 15 – Exemplo de parte de processo onde inúmeros pontos estão marcados como acesso restrito.


GRUPO CYRELA
A lista completa das empresas do Grupo Cyrela e suas respectivas atividades econômicas desempenhadas no Brasil, classificadas de acordo com a CNAE, integra esta notificação como DOCUMENTO DE ACESSO RESTRITO II.5.2.

Fonte: Adaptado da versão pública do ato de concentração 08700.001339/2021-81, página 6.

4.2.4.6.1 Inconsistência cruzada

A inconsistência cruzada acontece quando dado duas empresas **A** e **B**, **B** é listada no grupo econômico no ato de concentração de **A**, porém **A** não é listada no grupo econômico do ato de concentração de **B**. Em teoria isso demonstra uma inconsistência que deve ser analisada. Ao submeter o algoritmo à construção dos grupos econômicos de **A** ou **B** os valores serão diferentes e haverá uma penalização em falsos-positivos ou negativos devido a uma inconsistência de preenchimento do formulário de notificação.

4.2.4.6.2 Inconsistência sequencial

A inconsistência sequencial advém de que as mesmas empresas ao solicitarem atos de concentração seguidos, apresentam grupos econômicos distintos do que foi apresentado. Isso se justificaria em caso de saída de participação, desindustrialização, entre outros casos, porém, o tempo transcorrido sucessivas solicitações torna essas hipóteses pouco prováveis. *Por que não há consistência nas informações dos grupos econômicos apresentados ao longo dos processos? Quais as gravidades dessas inconsistências para a avaliação do casos?* Essas são perguntas que devem ser analisadas com cautela em próximos trabalhos.

4.2.4.6.3 Inconsistência de situação cadastral

A inconsistência de situação cadastral é uma situação em que empresas em situação baixada, nula ou suspensa são listadas como pertencente do grupo econômico. Nessa situação, o algoritmo não irá incluir essas empresas, por filtrar unicamente por empresas ativas, e, infelizmente será penalizado. A empresa corretamente não identificada aparecerá no falso negativo da medição, prejudicando os resultados. Infelizmente, os formulários não possuem

nenhuma etapa de verificação de inconsistências.

4.2.5 Métricas

Para validação deste trabalho foram utilizadas duas métricas oriundas da recuperação de informação: precisão e revocação (BUCKLAND; GEY, 1994). Dessa forma, sejam:

s : o conjunto de empresas selecionadas pelo algoritmo desenvolvido para compor o grupo econômico;

c : o subconjunto de s contendo as empresas que realmente fazem parte do grupo econômico. Isto é, excluindo-se os falsos positivos;

t : o conjunto de todas as empresas que fazem parte do grupo econômico informado via AC para uma determinada empresa de identificador *x*;

Na presente validação o conjunto utilizado como referência foi aquele informado via formulário de notificação.

4.2.6 Precisão

A precisão irá medir a proporção entre as empresas selecionadas corretamente em relação a tudo o que foi selecionado pelo algoritmo. "*De tudo o que selecionei, quanto é útil?*". Nesse caso, os falso positivos irão interferir na medida. Dessa forma:

p(x) : a precisão do algoritmo para o grupo econômico de uma empresa de identificador x pode ser definido como sendo:

$$p(x) = \frac{c}{s} \tag{4.1}$$

A revocação mede a proporção entre o que foi selecionado corretamente pelo algoritmo e tudo o que deveria ter sido selecionado. Isto é: "*De tudo o que deveria ser selecionado, quanto estava presente na solução?*". Dessa forma:

r(x) : a revocação medida para o algoritmo em relação ao grupo econômico de uma empresa de identificador x pode ser definida como sendo:

$$r(x) = \frac{c}{t} \tag{4.2}$$

Enquanto a precisão mede a utilidade do resultado, a revocação mede sua completude. Nos casos descritos, o erro tipo I, ou seja, o falso-positivo representa as empresas classificadas como membros do grupo econômico, quando na verdade não são. Por outro lado, o falso-negativo (erro tipo II) representa as empresas que fazem parte do grupo econômico mas o algoritmo não foi

capaz de relacionar.

Erros do tipo I podem ocorrer quando:

- A empresa informou à receita uma configuração de relação societária diferente da informada ao Cade;
- O percentual de participação da empresa é inferior ao mínimo previsto em resolução para constar de forma obrigatória no AC, sendo este o motivo da não inclusão da empresa no formulário de notificação;

Erros do tipo II podem ocorrer quando:

- A base pública da Receita Federal sendo consultada está trabalhando com um valor desatualizado, uma vez que as atualizações são realizadas a cada 3 meses;
- Empresa ou *holding* estrangeira sem **cnpj** no Brasil, que apesar de listadas nos formulários de notificação não existem na base pública da Receita Federal e sempre são contabilizadas como erros.

4.3 Resultados e Discussão

Cada empresa presente no *benchmark* teve o seu **cnpj** (identificador) submetido ao algoritmo desenvolvido. Em seguida, foram comparados o grupo econômico retornado pelo algoritmo com o grupo de empresas esperado presente no *benchmark* e a partir daí foram extraídas as métricas de precisão e revocação para todos os casos. Os resultados de cada execução são apresentados na Figura 17 onde cada ponto do gráfico representa um par ordenado (precisão, revocação). A Figura 16 apresenta parte da tabela[9] que contém o resultado da medição da revocação e precisão e erros para cada processo contido no *benchmark*.

Com a execução foram obtidos os seguintes resultados:

- Precisão média de 71% e revocação média de 48% como medição piso (quanto maior melhor em ambas as métricas). Isto é, as medições exploraram sempre o pior cenário proposto, podendo estas medidas serem melhores do que o previsto caso sejam implementadas as devidas correções para os problemas levantados no item 4.2.4;

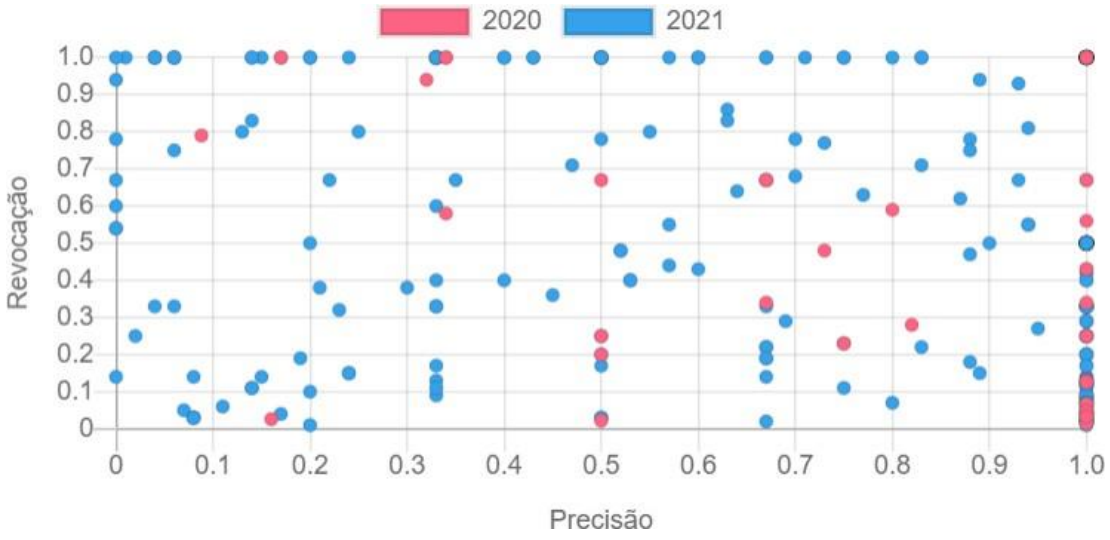
[9] É possível solicitar acesso completo aos dados da tabela entrando em contanto com o autor e realizando à solicitação ao Cade.

Figura 16 – Gráfico precisão vs revocação para cada uma das empresas que solicitaram AC em junho de 2020 e maio de 2021.

	Processo	A.	Empresa	CNPJ	Prec...	Rev...	Match	Erro - I	Erro - II
1	08700.000127/2021-86	2021	Grupo Danaher	13539558	0.63	0.86	-CYTIVA DO BRASIL COME...	-MOLECUL...	-Biosafe do B...
2	08700.000128/2021-21	2021	Grupo Econômico Geribá	10467534	1.00	0.13	-Grupo Econômico Geribá(...		-Winkel Cons...
3	08700.000128/2021-21	2021	Concessionária Rodovias d...	10678505	1.00	0.50	-Concessionária Rodovias ...		-AB Concessõ...
4	08700.000129/2021-75	2021	Columbia TriStar Filmes do...	979601	0.67	0.14	-COLUMBIA TRISTAR FILME...	-BUENA VIS...	-Sony Picture...
5	08700.000129/2021-75	2021	Warner Bros. Inc. Brazil Bra...	8437475	1.00	0.09	-Warner Bros. Inc. Brazil Br...		-Brasil Chann...
6	08700.000149/2021-46	2021	Grupo Localiza	16670085	1.00	0.13	-Grupo Localiza(16670085)		-Rental Brasil ...
7	08700.000149/2021-46	2021	Unidas S.A.	4437534	1.00	0.04	-Unidas S.A.(4437534)		-SF 166 Partic...
	08700.000150/2021-71	2021	Barra Energia do Brasil Pet...	9589793	0.33	1.00	-BARRA ENERGIA DO BRAS...	-FR BARRA ...	

Fonte: Elaboração própria.

Figura 17 – Gráfico precisão vs revocação para cada uma das empresas que solicitaram AC em junho de 2020 e maio de 2021.



Fonte: Elaboração própria.

- 100% dos falsos-positivos encontrados, apesar de não estarem registrados nos formulários de notificação, realmente possuíam um caminho societário na base da Receita Federal;
- 90% dos formulários de notificação analisados possuíam alguma inconsistência. Sejam elas pela situação cadastral das empresas citadas (empresas baixadas, inativas ou suspensas), problemas ortográficos que impedem a identificação da empresa sendo citada, empresas homônimas que levam a **cnpjs** distintos, incoerência entre as informações prestadas ao Cade e a RECEITA, entre outros problemas foram comumente

encontrados na análise;

- A obrigatoriedade de um identificador de empresa tal como o **cnpj** mostrou-se de grande importância para todas as análises dessa natureza.

5. DESENVOLVIMENTO DA FERRAMENTA BIMO

De modo a tornar o projeto de construção automática de grupos econômicos útil aos membros do Cade que necessitem dessa informação em alguma esfera do seu trabalho, decidiu-se por construir uma ferramenta, e disponibilizá-la na rede interna da instituição. Dentre os principais propósitos no desenvolvimento dessa ferramenta, encontram-se:

- Validar os resultados apresentados pela ferramenta aos membros do Cade com expertise no tema de grupos econômicos;
- Medir o nível de utilidade do algoritmo desenvolvido para a construção de grupos econômicos na compreensão de mercados relevantes;
- Verificar se as informações prestadas pelas empresas à Receita Federal são suficientes para o cumprimento do propósito de construção automática de grupos econômicos à luz das sugestões recebidas.

A ferramenta encontra-se disponível na rede interna do Cade por meio do seguinte link: http://bimo.cade.gov.br/bimo/.

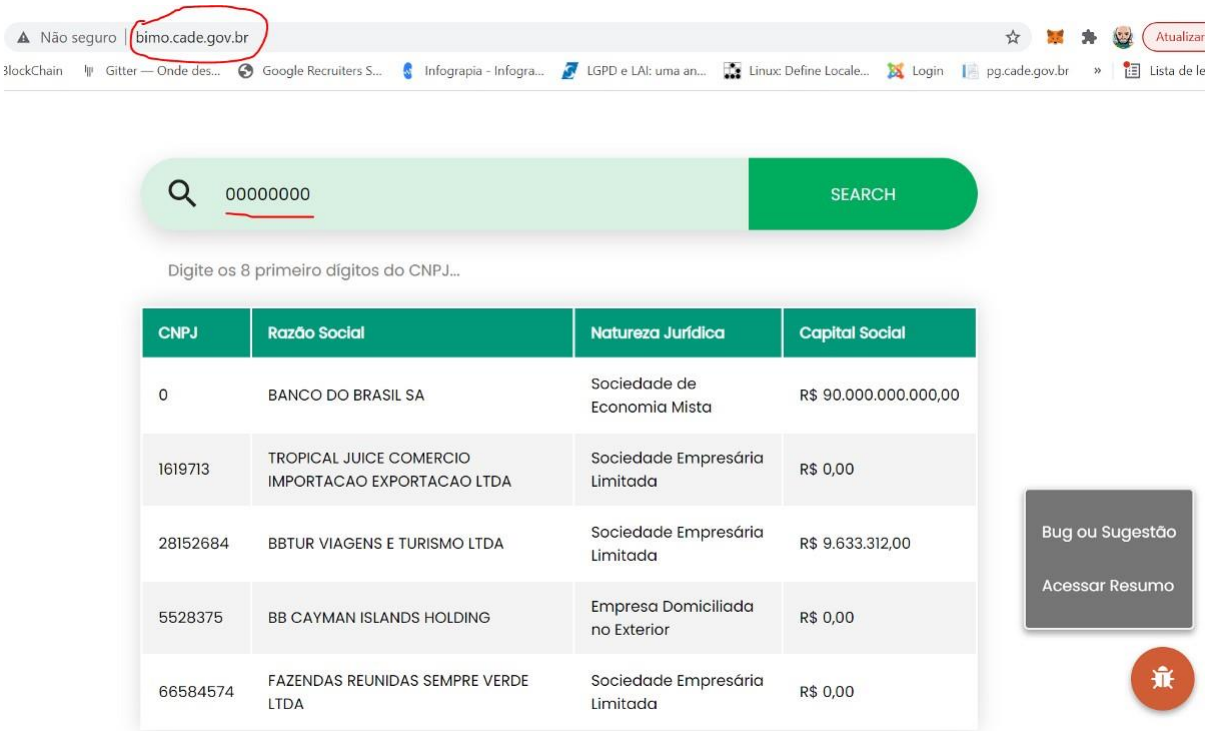
5.1 Interface gráfica

A interface gráfica do projeto foi desenvolvida para proporcionar a utilização mais simples e intuitiva possível. A Figura 18 apresenta uma ilustração da ferramenta Bimo em funcionamento na rede interna do Cade.

As principais considerações para a construção da interface foram:

1. Fundo branco e sem distrações;
2. Barra de consulta, semelhante à do *google*, para transmitir uma sensação de naturalidade e fluidez;
3. Inibição de letras e símbolos da barra de consulta. Apenas números são permitidos para digitar. Isso minimiza a possibilidade do usuário realizar solicitações que não sejam cnpjs e também elimina a sobrecarga de solicitações inúteis;

Figura 18 – Ilustração de funcionamento do bimo na rede interna do Cade.



Fonte: Elaboração própria.

4. Digitação de apenas 8 dígitos, referentes aos 8 primeiros dígitos do cnpj não sendo necessário digitar o cnpj completo. Gerando otimização do processo de consulta;
5. Permissão para copiar/colar cnpjs consultados de páginas da internet de modo a permitir o máximo de conforto em consultas;
6. Relatório em formato tabular que pode ser copiado para outras ferramentas e realizado download.

5.2 Desenvolvimento de uma API

A ferramenta foi construída em formato de API REST (MASSE, 2011) para que o algoritmo utilizado nas consulta pudesse ser utilizado em outros projetos desvinculados da interface sugerida, seja de modo a complementar alguma análise ou relatório externo, podendo receber múltiplas requisições automatizadas por meio do seguinte ponto de acesso:

Requisição: http://bimo.cade.gov.br/search/{cnpj}

O acesso via API permite que multiplas requisições possam ser realizadas sem a necessidade de interface de modo a alimentar algum sistema externo que necessidade de informações de grupos econômicos em suas análises.

5.3 Manutenibilidade

Esta ferramenta, de forma geral, foi projetada para ter o mínimo de interferência dos desenvolvedores que se proponham a continuá-lo. O processo de desenvolvimento é enxuto, baseado no princípio de integração contínua: fazer-enviar-publicar. Tão rápida uma melhoria é feita, ela é imediatamente disponibilizada ao usuário final. Neste sentido há:

1. Script de *boot* projetado para ser executado assim que a VM for iniciada. No sistema operacional Centos 7, presente na VM (10.1.x.xx), a construção dessa funcionalidade só foi possível após a construção de um serviço que foi denominado **bimo.service**. Um exemplo de como esse serviço foi construído encontra-se no repositório interno de acesso restrito aos funcionários;

2. Integração contínua com o *gitlab* que permite que as novas implementações e melhorias entrem em produção sempre que o desenvolvedor acione o *git pull* e reinicie o serviço.

5.4 Limitações

A seguir são apresentadas algumas limitações deste trabalho.

5.4.1 Sobre a construção ideal de grupos econômicos

A principal limitação deste trabalho é que o seu pilar de sustentação da validação tem como referência principal as informações prestadas pelas empresas nos formulários de notificação, o que não necessariamente representa a verdade. E se por algum motivo a empresa decidiu omitir ao Cade ou à Receita Federal alguma relação societária? Resultaria que a comparação sendo realizada estaria sendo enviesada de acordo com as informações fornecidas pelas empresas, mas é possível que ao menos algumas vezes exista ruído entre o que é informado é o que deveria ser o ideal. Assim sendo, o paradoxo existente é de ainda que se estivesse construindo um algoritmo que buscasse construir grupos econômicos automaticamente, a forma como os resultados são validados condiciona-o a antever o que seria informado pelas empresas ao invés de em um algoritmo que se aproximasse do grupos econômico real. E se assim o fosse, as métricas de precisão e revocação bem como todas as conclusões deveriam ser sobre a capacidade em antever ou prever o conjunto de empresas que serão informadas e não necessariamente os grupo econômicos ideais. Assim sendo, não há como verificar, sem acesso às bases das juntas comerciais se os resultados produzidos e os grupos informados nos ACs analisados estão efetivamente corretos.

5.4.2 Sobre a caducidade dos resultados

Os resultados obtidos estão restritos aos processos analisados entre 2020 e 2021. Isto é, não é possível afirmar nada sobre o passado ou sobre o futuro além de uma estatística que foi desenvolvida por meio da amostra de dados trabalhada. É possível que em alguns cenários a base da receita disponível não acompanhe a situação cadastral das empresas e essa incompatibilidade momentânea gere ruído nos resultados extraídos. É interessante saber a quantidade de empresas atualizadas a cada trimestre para que esses valores possam ser estimados estatisticamente. Após isso, a base de dados disponível perde a validade e a execução do algoritmo sob a base desatualizada pode aumentar consideravelmente a quantidade de erros.

5.5 Reprodutibilidade

Todo o trabalho, códigos, scripts e demais implementações estão disponíveis para serem acessadas, executadas e reproduzidas internamente para os funcionários da autarquia, sendo de acesso restrito ao público externo.

6. CONCLUSÃO

A construção automática de grupos econômicos proposta por este trabalho apresentou desempenho relevante. O algoritmo proposto apresenta-se como uma alternativa para a verificação dos grupos econômicos fornecidos pelas empresas, construção automática de mercado relevante ou simulação de aplicações de multas.

Este trabalho também apresenta potencial para ser expandido a empresas que necessitem verificar relações societárias para tomarem decisões, como é o caso dos grandes conglomerados bancários que necessitam conhecer relações societárias entre empresas para decidirem quanto a realização de empréstimos e por vezes possuem apenas as informações do cartão CNPJ para realizarem essa tarefa.

6.1 Trabalhos Futuros

A seguir são apresentadas algumas propostas de trabalhos futuros que pos- sam emergir a partir do que foi realizado:

1. Hierarquização de grupos econômicos por CNAEs e subCNAEs. Isto é, a adaptação do algoritmo desenvolvido para que a construção de grupos econômicos leve em consideração apenas empresas que possuem mesmo CNAE e/ou subCNAE da empresa alvo sendo analisada. Uma vez que a maioria dos ACs envolvem empre- sas do mesmo

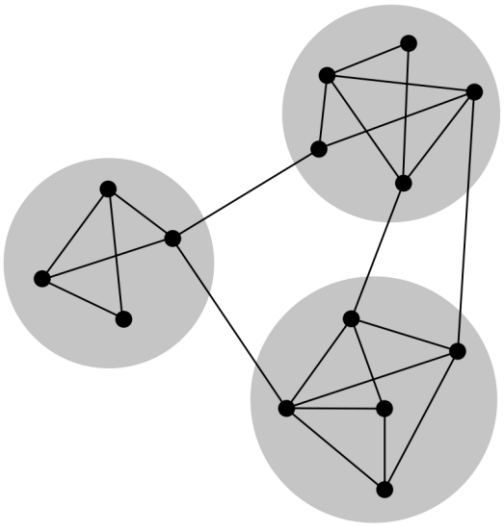
setor, a especialização prévia dos grupos apresenta potencial para simulações de HHI o que melhoraria a eficiência das análises necessárias na mai- oria dos casos;

2. Análise de inconsistências em formulários de notificação, passando a solicitar: cnpj para as empresas citadas nos grupos econômicos e sua verificação de status na base da Receita Federal. Isto é, com o formulário eletrônico que está sendo desenvolvido pelo Cade, uma etapa poderia ser a verificação direta entre as informações prestadas à Receita e as sendo prestadas ao Cade, eliminando as inconsistências já no momento de submissão do relatório;

3. Análise específica de relações societárias entre consórcio e *joint venture*. Atualmente existem ∼12.000 consórcios ativos na base da Receita Federal com múltiplas conexões entre grandes empresas. Conhecer mais profundamente este ambiente pode oportunizar grandes descobertas sobre construção de grupos econômicos;

4. Desenvolvimento de uma nova versão do algoritmo levando-se em consideração uma uma proposta de cluster e análise de grupos. A Figura 19 apresenta um exemplo de grupos bem definidos que sofrem coalescência em um único grupo, influ- enciados por relações simples entre empresas de grupos distintos;

Figura 19 – Exemplo de grupos bem definidos que foram agrupados em um único grupo por relações societárias simples.



Fonte: (ROY, 2022)

5. Construção de grupos econômicos levando-se em consideração as relações societárias entre pessoas físicas presentes na base da Receita Federal.

6. Monitoramento automático de atualização da base da Receita Federal.

REFERÊNCIAS

ALMEIDA, A. Economia aplicada para gestores. *Vila Nova de Gaia: Espaço Atlântico-Publicações e Marketing, Lda*, 2007. Citado na página 11.

BELLMAN, R. On a routing problem. *Quarterly of applied mathematics*, v. 16, n. 1, p. 87–90, 1958. Citado na página 25.

BORŮVKA, O. O jistém problému minimálním. 1926. Citado na página 25. BUCKLAND, M.; GEY, F. The relationship between recall and precision. *Journal of the*

American society for information science, Wiley Online Library, v. 45, n. 1, p. 12–19, 1994. Citado na página 40.

DIJKSTRA, E. W. et al. A note on two problems in connexion with graphs. *Numerische mathematik*, v. 1, n. 1, p. 269–271, 1959. Citado na página 25.

EISENTRAUT, P.; HELMLE, B. *PostgreSQL-administration*. [S.l.]: O'Reilly Germany, 2013. Citado na página 19.

FONTAINE, D. *The Art of PostgreSQL*. [S.l.]: First Edition, 2020. Citado na página 19.

FRONTIER, U. N. Exploiting sparc buffer overflow vulnerabilities. In:___. [s.n.], 2022. Disponível em: <http://www.ouah.org/UNF-sparc-overflow.html>. Citado na página 27.

GARCIA-MOLINA, H.; SALEM, K. Main memory database systems: An overview. *IEEE Transactions on knowledge and data engineering*, IEEE, v. 4, n. 6, p. 509–516, 1992. Citado na página 29.

HAGBERG, A.; SWART, P.; CHULT, D. S. *Exploring network structure, dynamics, and function using NetworkX*. [S.l.], 2008. Citado na página 29.

HELMER, O. *Analysis of the future: The Delphi method*. [S.l.], 1967. Citado na página 36. JATANA, N. et al. A survey and comparison of relational and non-relational database.

International Journal of Engineering Research & Technology, Citeseer, v. 1, n. 6, p. 1–5, 2012. Citado na página 18.

JI, S. Prim's algorithm. In:__. [s.n.], 2022. Disponível em: <https://en.wikipedia.org/wiki/Prim27s_algorithm#/media/File:PrimAlgDemo.gif>. Citado na página 25.

MASSE, M. *REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces*. [S.l.]: "O'Reilly Media, Inc.", 2011. Citado na página 45.

PIMENTA, G. Exploiting sparc buffer overflow vulnerabilities. In:___. [s.n.], 2019. Disponível em:<https://www.jota.info/coberturas-especiais/inova-e-acao/projeto-cerebro-cade-usa inteligencia-artificial-no-combate-a-carteis-29102019>.Citadonapágina15.

PLATTNER, H. et al. *A course in in-memory data management*. [S.l.]: Springer, 2013. Citado na página 29.

PRIM, R. C. Shortest connection networks and some generalizations. *The Bell System Technical Journal*, Nokia Bell Labs, v. 36, n. 6, p. 1389–1401, 1957. Citado na página 25.

QUINN, A. J.; BEDERSON, B. B. Human computation: a survey and taxonomy of a growing field. In: *Proceedings of the SIGCHI conference on human factors in computing systems*. [S.l.: s.n.], 2011. p. 1403–1412. Citado na página 36.

ROUSE, S. V. A reliability analysis of mechanical turk data. *Computers in Human Behavior*, Elsevier, v. 43, p. 304–307, 2015. Citado na página 36.

ROY, S. Clustering graph. In:____. [s.n.], 2022. Disponível em: <https://www.sciencedirect.com/topics/computer-science/clustering-graph>. Citado na página 49.

TOTH, A. Busca em largura. In:___. [s.n.], 2022. Disponível em: <https://pt.wikipedia.org/wiki/Busca_em_largura#/media/Ficheiro:6n-graf.svg>. Citado na página 26.

YWORKS. Edge betweeness clustering. In:___. [s.n.], 2022. Disponível em: <https://www.yworks.com/pages/clustering-graphs-and-networks>. Citado na página 28.