

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Processamento de Dados do Comércio Internacional
Usando o Modelo de Grafos do Neo4j

Clayton Miccas Junior



São Carlos – SP

Processamento de Dados do Comércio Internacional Usando o Modelo de Grafos do Neo4j

Clayton Miccas Junior

***Orientador:* Profa. Dra. Elaine Parros Machado de Sousa**

Monografia final de conclusão de curso apresentada
ao Instituto de Ciências Matemáticas e de
Computação – ICMC-USP, como requisito parcial
para obtenção do título de Bacharel em Engenharia
de Computação.

Área de Concentração: Banco de Dados

USP – São Carlos
Novembro de 2021

Junior, Clayton Miccas

Processamento de Dados do Comércio Internacional
Usando o Modelo de Grafos do Neo4j / Clayton
Miccas Junior. - São Carlos - SP, 2021.

52 p.; 29,7 cm.

Orientador: Elaine Parros Machado de Sousa.

Monografia (Graduação) - Instituto de Ciências
Matemáticas e de Computação (ICMC/USP), São Carlos -
SP, 2021.

1. Banco de dados. 2. Comércio internacional.
3. UN Comtrade. 4. Neo4j. 5. Grafos. I. Sousa,
Elaine Parros Machado de. II. Instituto de Ciências
Matemáticas e de Computação (ICMC/USP). III. Título.

Este trabalho é dedicado aos meus familiares, amigos e educadores, que sempre contribuíram com seus saberes e experiências para o meu desenvolvimento.

AGRADECIMENTOS

Os agradecimentos principais são direcionados à minha orientadora, Prof^a Dr^a Elaine Parros Machado de Sousa, e a seu aluno orientando de doutorado, Marcus Araujo. Ambos foram fundamentais para o desenvolvimento deste projeto, e nos momentos mais desafiadores, se dispuseram a me auxiliar.

Agradeço aos meus familiares, amigos, e, principalmente, meus pais, pois me deram suporte nos momentos mais complexos e souberam compreender as demandas que um projeto como este gera a uma pessoa.

*“We can only see a short distance ahead,
but we can see plenty there that needs to be done.”
(Alan Turing)*

RESUMO

JUNIOR, C. M.. **Processamento de Dados do Comércio Internacional Usando o Modelo de Grafos do Neo4j**. 2021. 52 f. Monografia (Graduação) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Este trabalho trata do desenvolvimento de um banco de dados de grafos, incluindo técnicas e estudos sobre as medidas usuais deste modelo de base de dados e sua respectiva performance. O trabalho aborda registros de transações de *commodities* entre os países e grupos econômicos presentes no repositório de dados *UN Comtrade*, possuindo os registros a granularidade de ano, país e produto comercializado, sendo que este é expressado através do índice de classificação de produtos HS. A construção da base de dados utilizando o *Neo4j Graph Platform* possibilitou o cálculo eficiente da medida de centralidade *PageRank* e a detecção de comunidades através do método *Louvain*, além da geração de visualizações acerca dos países e suas transições durante os anos. Os resultados obtidos mostram a grande performance que bancos de dados de grafos têm em relação a grandes massas de dados e sua utilização para a visualização de registros temporais.

Palavras-chave: Banco de dados, Comércio internacional, UN Comtrade, Neo4j, Grafos.

ABSTRACT

JUNIOR, C. M.. **Processamento de Dados do Comércio Internacional Usando o Modelo de Grafos do Neo4j**. 2021. 52 f. Monografia (Graduação) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

This work deals with the development of a database of graphs, covering techniques and studies on the present measures of this database model and its respective performance. The work addresses records of commodity transactions between countries and economic groups present in the *UN Comtrade* data repository, with the records having the granularity of year, country and traded product, which is expressed through the classification index of HS products. The construction of the database with Neo4j graph Platform enabled the development of a measure of centrality *PageRank* and the detection of communities through the *Louvain* method, thus generating views about the countries and their transitions over the years. The constructed results demonstrate the great performance that graph databases have in relation to large amounts of data and their use for visualizing temporal records.

Key-words: Database, International trade, Comtrade, Neo4j, Graph.

LISTA DE ILUSTRAÇÕES

Figura 1 – Grafo relacional entre duas pessoas e seu veículo	26
Figura 2 – Tipos de grafos	26
Figura 3 – Conectividade entre grafos	27
Figura 4 – Direcionamento de relacionamentos de grafos	27
Figura 5 – Peso de relacionamentos de grafos	28
Figura 6 – Apresentação de ciclos fechados em grafos	29
Figura 7 – Método Louvain: primeira iteração	30
Figura 8 – Ecossistema Neo4j Graph Platform	30
Figura 9 – Inserção de dados no UN Comtrade Database nos últimos 10 anos	33
Figura 10 – Exemplo de modelagem em grafos de transações comerciais entre países	35
Figura 11 – Subgrafos desenvolvidos para as transações entre países analisadas	36
Figura 12 – Grafos representativos dos anos de 2007 e 2008, respectivamente - centralidade (PageRank)	37
Figura 13 – Nuvem de palavras: países com maior significância de exportações nos anos de 2007 e 2008.	38
Figura 14 – Grafos representativos dos anos de 2007 e 2008 - detecção de comunidades (método Louvain)	39
Figura 15 – Mapa de visualização de comunidades nos anos de 2007 e 2008	39
Figura 16 – Diagrama de Sankey - Distribuição das comunidades	40
Figura 17 – Diagrama de Sankey - Distribuição das comunidades no ano de 2008	40
Figura 18 – Fluxos de transações dos países que se mantiveram na comunidade A	41
Figura 19 – Fluxos de transações dos países que se deslocaram na comunidade A	41
Figura 20 – Fluxos de transações dos países que se mantiveram na comunidade B	51
Figura 21 – Fluxos de transações dos países que se deslocaram na comunidade B	51
Figura 22 – Fluxos de transações dos países que se mantiveram na comunidade C	52
Figura 23 – Fluxos de transações dos países que se deslocaram na comunidade C	52

LISTA DE TABELAS

Tabela 1 – Lista de campos presentes na base de dados	34
Tabela 2 – Países mais bem posicionados em 2007 e 2008 pelo algoritmo PageRank . .	37

LISTA DE CÓDIGOS-FONTE

Código-fonte 1 – Cypher - Inserção de dados no Neo4j Graph Platform	35
Código-fonte 2 – Cypher - PageRank	37
Código-fonte 3 – Cypher - Louvain	38
Código-fonte 4 – Cypher - Transações e Comunidades	40

LISTA DE ABREVIATURAS E SIGLAS

UNSD United Nations Statistics Division

SUMÁRIO

1	INTRODUÇÃO	23
1.1	Motivação e Contextualização	23
1.2	Objetivos	24
1.3	Organização	24
2	MÉTODOS, TÉCNICAS E TECNOLOGIAS UTILIZADAS	25
2.1	Bases de dados de grafos	25
2.2	Centralidade e o Algoritmo PageRank	28
2.3	Detecção de Comunidades e o Método Louvain	29
2.4	Neo4j Graph Platform	30
3	DESENVOLVIMENTO	33
3.1	Considerações Iniciais	33
3.2	Atividades Realizadas	34
3.2.1	<i>Modelagem da base de dados</i>	<i>34</i>
3.2.2	<i>Criação da base de dados e inserção dos registros</i>	<i>35</i>
3.2.3	<i>Centralidade e o algoritmo PageRank</i>	<i>36</i>
3.2.4	<i>Detecção de comunidades e o método Louvain</i>	<i>38</i>
3.3	Resultados	42
3.4	Dificuldades e Limitações	43
4	CONCLUSÃO	45
4.1	Contribuições	45
	REFERÊNCIAS	47
ANEXO A	FLUXOS DE TRANSAÇÕES DE EXPORTAÇÕES, IMPORTAÇÕES E BALANÇAS COMERCIAIS NAS COMUNIDADES B E C	51

INTRODUÇÃO

1.1 Motivação e Contextualização

A economia mundial atual é em grande parte composta pelo comércio de *commodities*, definição que é atribuída a produtos primários com pequeno grau de industrialização (TOPIK; MARICHAL; FRANK, 2006). As relações bilaterais de comércio entre países desenvolveram-se intensamente no último século, e de acordo com a *United Nations Conference on Trade and Development*, o valor das exportações de *commodities* entre países aumentou em uma escala de cinco vezes entre os anos de 1995 e 2011 (United Nations Conference on Trade and Development, 2013).

De acordo com a *United Nations Statistics Division (UNSD)*, existem aproximadamente 5300 descrições de produtos e subprodutos na convenção de classificação de produtos descrita como *Harmonized Commodity Description and Coding Systems (HS)* (United Nations Statistics Division, 2018). O relatório *Trade data processing and validation methodology, 2019*, desenvolvido pela UNSD, traz que a *UN Comtrade* é um repositório que une, desde 1962, mais de 40 bilhões de registros (United Nations Statistics Division, 2020), sendo este o repositório com maior uso para estatísticas de comércio internacional de *commodities*, onde apresenta registros de importação e exportação de *commodities*, classificadas pelo índice HS, entre países descritos pela ISO 3166-1 Alpha-3 (United Nations Statistics Division, 2019).

A busca para compreender os efeitos que um determinado evento traz na economia mundial é contínua, tanto para se entender as causas de eventos passados como para se prever ocorrências e possíveis desfechos que uma determinada ocorrência traz a um país, ou grupo de países que se assemelham por determinada característica. Tomemos como exemplo o caso da crise financeira de 2008, causada pela concessão de empréstimos hipotecários no mercado norte-americano, a qual resultou em uma crise sistêmica que atingiu o mundo todo e pode ser considerada a maior crise econômica desde 1929 (BRESSER-PEREIRA, 2009).

No cenário atual do comércio internacional, especialistas de diferentes áreas têm à disposição quantidades massivas de dados gerados a cada ano, alimentando bases históricas e temporais como a da *UN Comtrade*. A análise desses dados e obtenção de métricas de interesse para, por exemplo, apoio a tomadas de decisão, demanda técnicas computacionais eficientes, com baixa latência de processamento. Nesse contexto, o foco deste trabalho de conclusão de curso é propor soluções computacionais para tratamento e análise dos dados da *UN Comtrade*.

1.2 Objetivos

O objetivo geral deste trabalho foi implementar um solução baseada em grafos para uma análise inicial de relações comerciais entre países e grupos econômicos ao longo do tempo. Para tanto, foram estabelecidos alguns objetivos específicos: 1) representar os dados da *UN Comtrade* numa estrutura de grafos; 2) detectar comunidades que correlacionem os países através de suas relações comerciais de importação/exportação; 3) computar métricas de interesse, como a centralidade; 4) identificar a movimentação dos referidos países entre as comunidades ao longo dos anos, permitindo, por exemplo, verificar as movimentações em períodos críticos como a crise de 2008.

Além disso, considerando a grande massa de dados advindos da *UN Comtrade*, os objetivos específicos deste trabalho incluem a utilização de um sistema de banco de dados de grafos, em particular o Neo4j, para armazenamento dos dados, criação dos grafos ao longo do tempo, e cálculo eficiente de comunidades e de métricas de interesse.

1.3 Organização

Este trabalho é apresentado em 4 capítulos, onde o primeiro descreve as informações introdutórias deste projeto. No Capítulo 2: Métodos, Técnicas e Tecnologias Utilizadas, são apresentados os conceitos técnicos, teorias e tecnologias. No Capítulo 3: Desenvolvimento, são descritas as atividades realizadas ao longo do projeto, abordando em detalhes suas especificidades. No quarto e último capítulo, Conclusão, são apresentados e discutidos os resultados obtidos do desenvolvimento, correlacionando-os aos objetivos iniciais e os passos futuros de desenvolvimento.

MÉTODOS, TÉCNICAS E TECNOLOGIAS UTILIZADAS

A seguir serão apresentadas as técnicas relevantes ao desenvolvimento deste trabalho, a terminologia empregada na literatura e a fundamentação teórica necessária para o entendimento de bases de dados de grafos, de detecção de comunidades, medidas de centralidade e das medidas de interesse pertinentes ao comércio internacional.

2.1 Bases de dados de grafos

Segundo [ROBINSON; WEBBER; EIFREM 2015](#), a linha do tempo, em alto nível, de bases de dados pode ser traduzida em quatro grandes tecnologias que permeiam o período entre 1960 e a atualidade: Dados Hierárquicos, Relacionais, NoSQL e Grafos, sendo este último abordado com maior intensidade nos últimos anos.

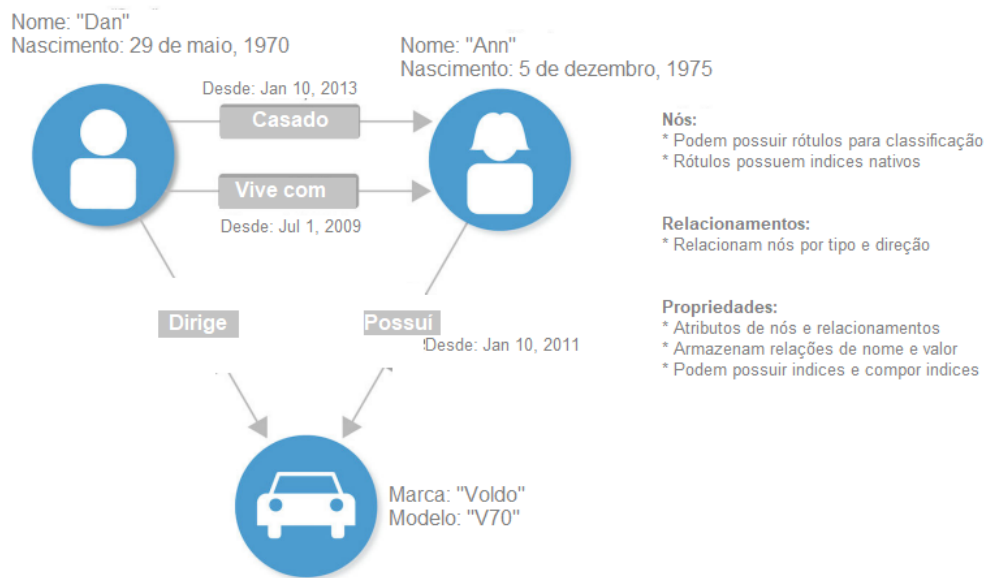
Bases de dados de Grafos possuem como uma de suas maiores atribuições a facilidade de se trabalhar com problemas complexos. Segundo [ROBINSON; WEBBER; EIFREM 2015](#), "*problemas complexos são problemas individuais que são observados e mensurados em sistemas complexos*". Já sistemas complexos, segundo os mesmos autores, se definem por: "*sistemas compostos por muitos componentes individuais que estão interconectados de várias maneiras, de modo que o comportamento geral não seja um único agregado dos componentes individuais*".

Uma representação gráfica de um grafo é apresentada na Figura 1, ilustrando por meio da relação entre duas pessoas e um veículo, as três principais características de um grafo: os nós, as relações e as propriedades. Um grafo é representado também por $G(V, T)$, onde V é o conjunto de nós (vértices) e T os pares relacionados (arestas).

Os nós são os elementos a serem analisados. Cada indivíduo daquele determinado sistema complexo é definido como um nó, que se relaciona com os demais elementos presentes no grafo através das arestas descritas como relações. Estes mesmos nós possuem características específicas a eles, definidas como propriedades. Uma propriedade pode demonstrar que um determinado nó pertence a um grupo de nós similares, por possuírem a mesma propriedade.

Grafos podem possuir tipos e estruturas ([NEEDHAM; HODLER, 2019](#)), sendo os três principais tipos: o grafo simples, em que um par de nós tem no máximo um relacionamento e não mais que um; os multi-grafos, que permitem que um determinado nó tenha mais que

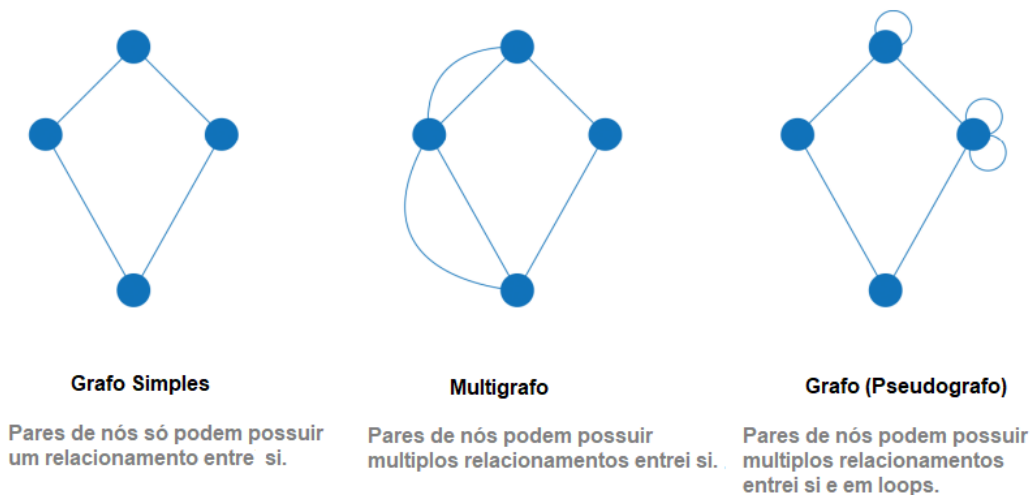
Figura 1 – Grafo relacional entre duas pessoas e seu veículo



Fonte: Traduzido de (NEEDHAM; HODLER, 2019)

um relacionamento; e, por fim, o pseudo grafo, que permite que nós tenham mais que um relacionamento e que possuam relacionamentos consigo mesmos. Essa tipificação é apresentada na Figura 2.

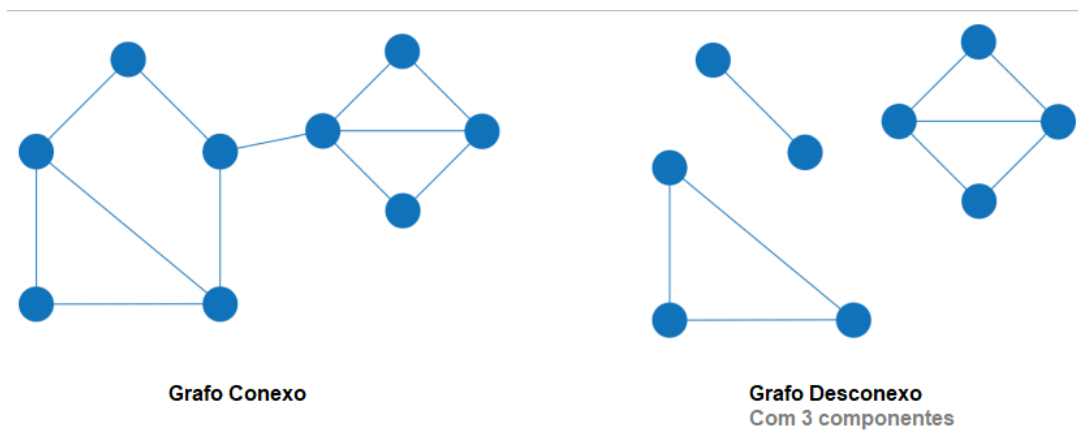
Figura 2 – Tipos de grafos



Fonte: Traduzido de (NEEDHAM; HODLER, 2019)

Estes tipos possuem algumas características gerais, como a conectividade entre os nós, onde um grafo pode ser caracterizado como conexo ou desconexo. Um grafo conexo é aquele em que existe pelo menos um caminho para se percorrer entre qualquer par de elementos. Esta tipificação é apresentada na Figura 3.

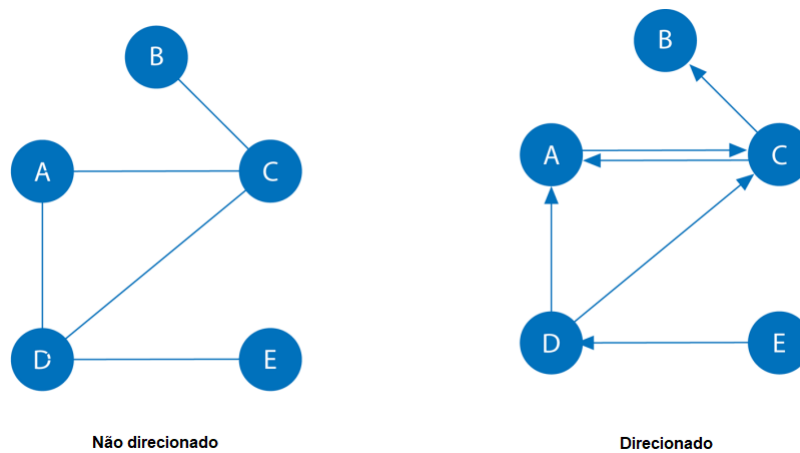
Figura 3 – Conectividade entre grafos



Fonte: Traduzido de ([NEEDHAM; HODLER, 2019](#))

Os relacionamentos entre nós possuem duas principais propriedades, sendo elas: direção e peso. Um relacionamento pode ser direcionado ou não direcionado, isso é, um relacionamento entre um nó "A" e um nó "B" é definido como não direcional se implicar em um relacionamento entre "B" e "A". Caso contrário o mesmo é caracterizado como um relacionamento direcionado, ilustrado na Figura 4.

Figura 4 – Direcionamento de relacionamentos de grafos



Fonte: Traduzido de ([NEEDHAM; HODLER, 2019](#))

Estes relacionamentos ainda podem possuir ou não peso, Figura 5, sendo o peso de um relacionamento o que descreve as métricas que o mesmo possui, tais como custo, tempo, distância, etc.

Figura 5 – Peso de relacionamentos de grafos



Fonte: Traduzido de (NEEDHAM; HODLER, 2019)

2.2 Centralidade e o Algoritmo PageRank

Segundo NEEDHAM; HODLER 2019, centralidade é sobre entender o quão importante é um nó em um determinado grafo, de modo a entender o impacto que um elemento traz em um grafo. Existem diversos modos para se configurar a importância de um elemento em um grafo e consequentemente existem diferentes maneiras de se calcular esta importância. ERCSEY-RAVASZ *et al.* 2012 traz referências sobre medidas de centralidade datadas de 1869, desenvolvidas pelo matemático Camille Jordan.

Há diversas medidas de centralidade, tais como: centralidade de grau, de proximidade, de intermediação, de auto vetor e baseada na dissimilaridades. O estudo da centralidade em grafos é fundamental para análises preliminares de um grafo, (BRANDÃO; DEL-VECCHIO; MELLO, 2015) e um dos mais recomendados algoritmos é o PageRank (NEEDHAM; HODLER, 2019), algoritmo esse notavelmente discutido no trabalho *The Anatomy of a Large-Scale Hypertextual Web Search Engine* (BRIN; PAGE, 1998) e, posteriormente, aplicado a um dos maiores motores de buscas desenvolvidos para internet, o *Google*.

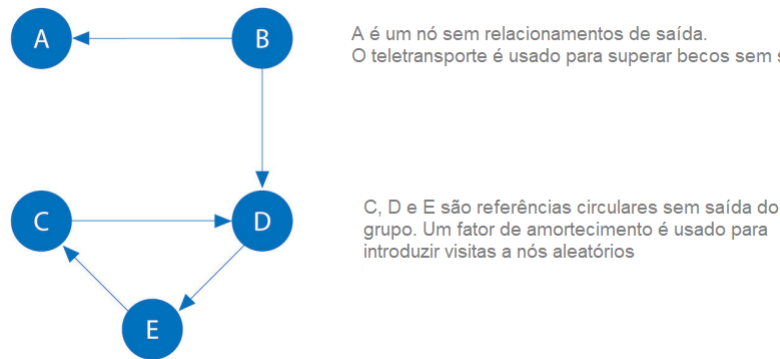
O valor PageRank dos "n" nós de um grafo em sua primeira iteração é definido como $\frac{1}{n}$. Para as demais iterações, até a convergência ou número pré determinado, o valor têm sua equação definida por NEEDHAM; HODLER 2019 :

Equação 1.

$$PR(\mu) = (1 - d) + d \left(\frac{PR(T1)}{C(T1)} + \dots + \frac{PR(Tn)}{C(Tn)} \right)$$

Onde μ é o nó para o qual se busca o valor do PageRank e que tem relações com os nós $T1$ a Tn , d varia entre 0 e 1 e é descrito como o fator de amortecimento (comumente se usando o valor de 0.85 (BOLDI; SANTINI; VIGNA, 2005)). Utilizado para se contornar a problemática do algoritmo entrar em ciclos fechados entre os nós, situação apresentada na Figura 6. A variável $C(tn)$ representa o grau de saída de um determinado nó, isso é, quantas relações partem deste nó.

Figura 6 – Apresentação de ciclos fechados em grafos



Fonte: Traduzido de (NEEDHAM; HODLER, 2019)

2.3 Detecção de Comunidades e o Método Louvain

Segundo NEEDHAM; HODLER 2019, a formação de comunidades é comum em todos os tipos de grafos, e a identificação destas é essencial para os estudos de grupos e seus fenômenos. De maneira geral, a detecção de comunidades tem como papel identificar os nós com relações mais similares entre si e posteriormente produzir e apresentar a visualização destas comunidades e suas respectivas métricas e características.

Assim como a centralidade, existem diversos algoritmos de detecção de comunidades, cada um destes com suas especificidades, tais como a necessidade do grafo possuir ou não pesos, ser direcional ou não, entre outras. Cita-se a contagem de triângulos, o coeficiente de clusterização, componentes conectados fortemente ou fracamente e o algoritmo de modularidade de Louvain (NEEDHAM; HODLER, 2019).

O algoritmo de modularidade de Louvain (BLONDEL *et al.*, 2008), busca a melhor otimização da modularidade dos nós de um grafo. Modularidade é uma medida proposta para se avaliar a qualidade das comunidades encontradas (GIRVAN; NEWMAN, 2015) e definida como (NEEDHAM; HODLER, 2019):

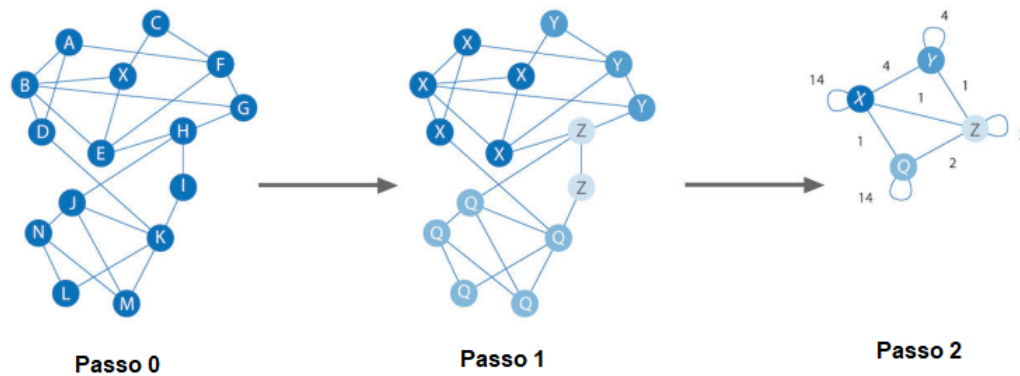
Equação 2.

$$M = \sum_{c=1}^{n_c} \left(\frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right)$$

Onde L é o número de relacionamentos de um grupo todo, L_c é o número de relacionamentos de uma determinada partição do grafo, e k_c se define como o grau dos nós daquela dada partição.

Após o cálculo da modularidade local dos nós, uma segunda etapa é realizada, a junção dos nós de uma mesma comunidade, assim definindo um novo grafo. Estas duas etapas são iteradas até que se tenha a otimização do valor para uma modularidade global das comunidades. A Figura 7 ilustra a primeira iteração do processo.

Figura 7 – Método Louvain: primeira iteração



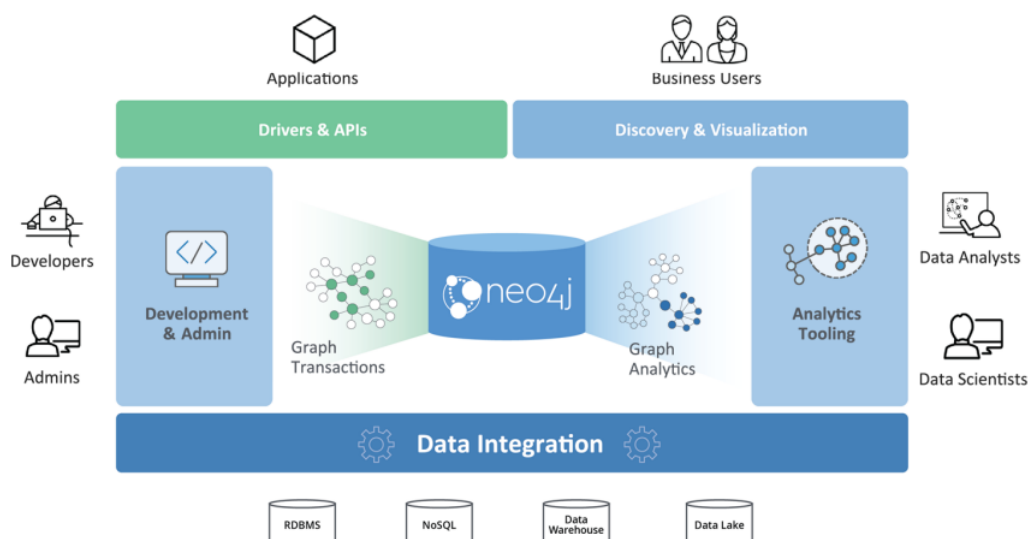
Fonte: Traduzido de (NEEDHAM; HODLER, 2019)

2.4 Neo4j Graph Platform

*Neo4J Graph Platform*¹ é um provedor de tecnologia de grafos construído em Java, com ferramentas de acesso públicas e privadas. Segundo [Neo4j company 2020](#), é a ferramenta de tecnologia de grafos utilizada em 75% das empresas listadas na Fortune 100 ([Fortune company, 2021](#)), índice norte americano que apresenta as maiores empresas do mundo.

Neo4J Graph Platform possui um sistema completo para o processamento de dados em grafos, como *bulks* para a integração de dados contidos em *data lakes* e *data warehouses*, ferramentas de processamento e análise de dados, e até ferramentas de visualização. A Figura 8 apresenta este sistema e seus usuários de interesse.

Figura 8 – Ecossistema Neo4j Graph Platform



Fonte: Needham e Hodler (2019).

¹ <<https://neo4j.com/developer/graph-platform/>>

A plataforma utilizada neste trabalho será a de uso público e as ferramentas utilizadas serão Neo4j *Bloom* e *Browser*. Esta última responsável pela introdução de códigos fontes em linguagem Cypher², linguagem inspirada em SQL e desenvolvida para a realização de *queries* no sistema. A título de comparação de sua performance com a ferramenta MySQL³, utilizando bases de dados compostas por 1 milhão de elementos, determinadas consultas percorrem a base de dados desenvolvida no Neo4j em 1 centésimo de tempo em relação à base desenvolvida em MySQL (VUKOTIC; WATT, 2014).

² <<https://neo4j.com/developer/cypher/>>

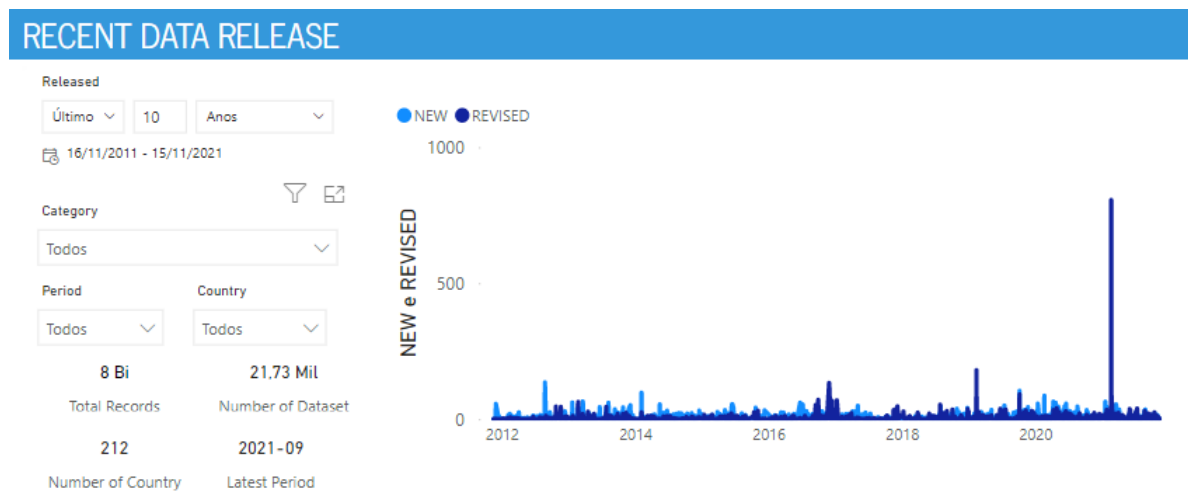
³ <<https://www.mysql.com/>>

DESENVOLVIMENTO

3.1 Considerações Iniciais

O repositório de dados *UN Comtrade* disponibiliza dados estruturados sobre as relações comerciais dos países e grupos econômicos desde 1962 ([United Nations Statistics Division, 2020](#)). O repositório recebe grande massa de dados com o passar do tempo, e, como apresenta a Figura 9, nos últimos 10 anos houve a inserção de mais de 8 bilhões de registros e mais de 21 mil bases de dados no sistema ([UN Comtrade Database, 2021](#)).

Figura 9 – Inserção de dados no UN Comtrade Database nos últimos 10 anos



Fonte: [UN Comtrade Database \(2021\)](#).

Dada a massividade de dados da *UN Comtrade* que a base apresenta, este trabalho traz como proposta a criação de um grafo e a otimização do tempo de consultas sobre estes dados, cálculos de medidas de centralidade e outras métricas de interesse, determinação de comunidades entre os elementos de interesse e suas movimentações dos países entre comunidades no decorrer do tempo amostral de 2000 a 2010, focando nas principais movimentações da crise de 2008.

Para esta proposição será apresentada a representação de grafos para os elementos estudados, suas aplicações, a ferramenta Neo4j e os desenvolvimentos necessários. Na sequência, serão apresentados os resultados obtidos e por fim as dificuldades, limitações e considerações finais acerca do projeto.

3.2 Atividades Realizadas

Esta seção apresenta em detalhes as etapas da desenvolvimento deste projeto, e os resultados obtidos.

3.2.1 Modelagem da base de dados

A base de dados trabalhada é composta por seis colunas de dados, sendo estas apresentadas na Tabela 1.

Tabela 1 – Lista de campos presentes na base de dados

Dados	Tipo	Descrição
year	inteiro	Ano da transação
origin	string	País de origem - ISO 3166, país exportador
dest	string	País de destino - ISO 3166, país importador
hs96	inteiro	Índice HS de produtos
export_val	float	Valor registrado de exportação pelo país de origem
import_val	float	Valor registrado de importação pelo país de origem

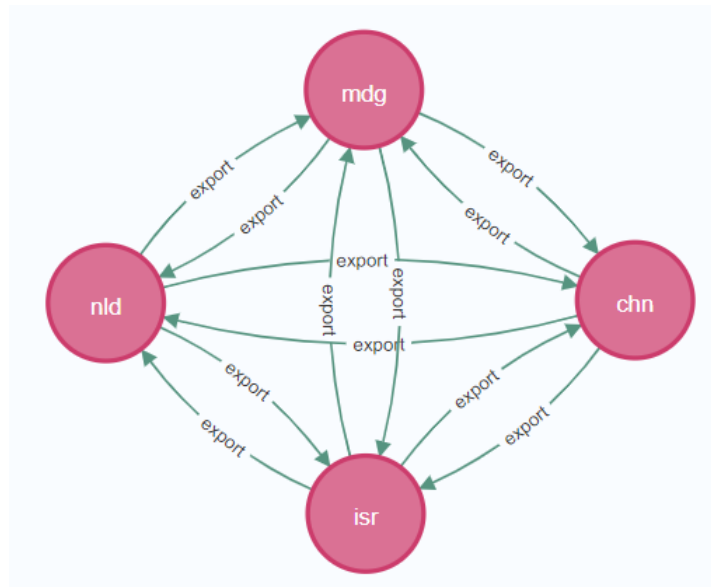
Fonte: [Desenvolvida pelo autor \(2021\)](#).

A modelagem proposta para a construção do grafo leva em consideração apenas quatro dos campos disponíveis. Valores do índice HS de produtos foram, a primeiro momento, agregados como um só produto, e a justificativa para agregação se dá devido ao grande aumento da complexidade da análise; Valores de importação foram removidos, devido a divergências entre os valores apresentados pelo exportador e pelo importador para um determinado produto. Essa diferença pode ser atribuída a diversos fatores, como erros cadastrais, divergência na data de registro e consequentemente alteração dos balanços anuais, entre outras.

Desta forma, o grafo foi construído tal que: todo país, de origem ou destino, é caracterizado como um nó. Os nós se relacionam através da existência de exportações naquele determinado ano, e para esta relação é atribuído como peso o valor registrado de exportações e o sentido da mesma é levado em consideração. Por fim, os nós deste grafo possuem como propriedade o ano em que as transações ocorreram, esta propriedade segmenta o grafo em sub-grafos, respectivos aos anos das transações. Foi construído, portanto, um multi grafo, desconexo, direcional e ponderado.

A Figura 10 ilustra parte deste grafo, com quatro nós e suas relações. Os valores de suas propriedades e o peso das relações não são apresentados visualmente, mas disponibilizados pela plataforma *Neo4j* em um sub menu lateral.

Figura 10 – Exemplo de modelagem em grafos de transações comerciais entre países



Fonte: Desenvolvida pelo autor (2021).

3.2.2 Criação da base de dados e inserção dos registros

Dada a configuração apresentada no tópico anterior, a criação da base de dados e do grafo se dá através do Código-fonte 1, apresentado abaixo. Esta codificação está apresentada em linguagem *Cypher*, onde em sua primeira linha é apontada a quantidade de elementos inseridos por vez ao sistema, posteriormente, na linha 2, o tipo e caminho do arquivo que contém os dados e, por fim, entre as linhas 3 e 9, a estruturação em si do grafo, onde se realizam os agrupamentos necessários de modo a se levar em conta todas as transações, mas não separando as mesmas pela codificação HS, diminuindo assim os números de relações.

Código-fonte 1: Cypher - Inserção de dados no Neo4j Graph Platform

```

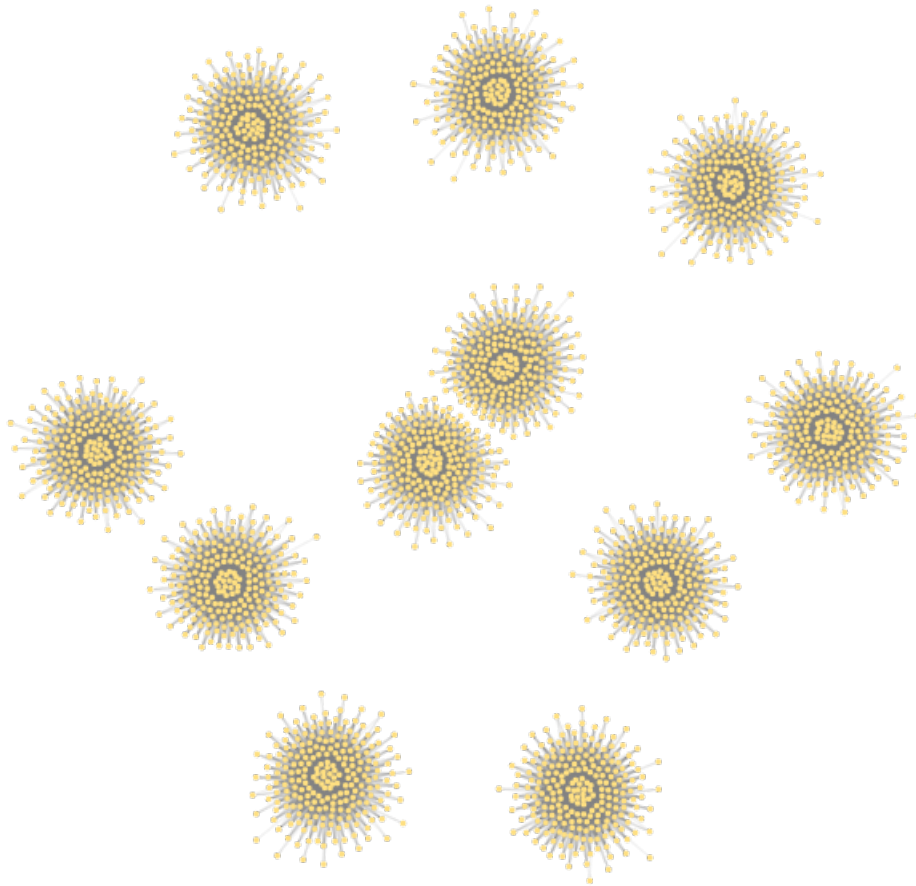
1 :auto USING PERIODIC COMMIT 500
2 LOAD CSV WITH HEADERS FROM "file:///dataset.csv" AS dados
3 WITH dados
4 WHERE dados.origin IS NOT NULL
5 MERGE (n:Pais {Name:dados.origin, Year: toInteger(dados.year)})
6 MERGE (m:Pais {Name:dados.dest, Year: toInteger(dados.year)})
7 MERGE (n) -[r:export] -> (m)
8 ON CREATE SET r.weight = coalesce(toInteger(dados.export_val), 0)
9 ON MATCH SET r.weight = r.weight + coalesce(toInteger(dados.
  export_val), 0)

```

A implementação e inserção dos dados resulta em um grafo com 2564 nós, 77148 relações e particionado em 11 subgrafos, com aproximadamente 233 nós por subgrafo. Como cada subgrafo representa as transações de um determinado ano, e a modelagem proposta de um

nó o descreve com a propriedade do ano, o mesmo país em anos distintos é apresentado por nós distintos. Esta construção está representada na Figura 11.

Figura 11 – Subgrafos desenvolvidos para as transações entre países analisadas



Fonte: [Desenvolvida pelo autor \(2021\)](#).

3.2.3 Centralidade e o algoritmo PageRank

Posteriormente à criação do grafo, foi computada a primeira medida de interesse, o cálculo da centralidade dos nós, a partir do algoritmo de *PageRank*. Será utilizado o valor empírico de 0.85, aplicado por (BRIN; PAGE, 1998), e o número máximo de iterações para convergência foi definido como 20, valores correspondentes aos definidos como padrões pelo *Neo4j*.

A aplicação do algoritmo para cada um dos nós presentes no grafo é desenvolvida através da já citada linguagem *Cypher*, e o Código-fonte 2 traduz essa construção na aplicação para cada um dos subgrafos. Após o cálculo da métrica, esta é atribuída ao seu respectivo nó como uma propriedade denominada *PageRank*, utilizada posteriormente na visualização dos países com maior centralidade.

Código-fonte 2: Cypher - PageRank

```

1 CALL gds.pageRank.write('Grafo_2008', {
2   maxIterations: 20,
3   dampingFactor: 0.85,
4   relationshipWeightProperty: 'weight',
5   writeProperty: 'pagerank'})

```

Os valores obtidos para o *PageRank* são aplicados como parâmetros de tamanho do nó na visualização do grafo. Assim, temos os países com maior centralidade representados com maior importância, com maiores registros de exportação e com relações mais significativas devido a lógica aplicada pelo algoritmo. A título de exemplo, são apresentados na Tabela 2 os países com maiores valores de *PageRank* nos anos de 2007 e 2008 e na Figura 12 suas representações no grafo, quanto maior a centralidade maior o tamanho do nó. Para comparação visual, a Figura 13 apresenta uma representação gráfica do tipo nuvens de palavras disponibilizada pelo *UN Comtrade*, a qual traz os países com as exportações mais significativas nos anos de 2007 e 2008 ([UN Comtrade Database, 2021](#)), diferente do algoritmo *PageRank* a nuvem de palavras é desenvolvida considerando apenas a somatória das exportações e não leva em consideração as relações para o cálculo.

Tabela 2 – Países mais bem posicionados em 2007 e 2008 pelo algoritmo PageRank

País	PageRank (Ano 2007)	PageRank (Ano 2008)
chn (China)	42.57	42.77
jpn (Japão)	27.09	26.05
kor (Coreia do Sul)	16.10	15.72
usa (Estados Unidos da América)	11.62	10.24
hkg (Hong Kong, China)	9.80	9.47
ind (Índia)	9.17	9.77

Fonte: [Desenvolvida pelo autor \(2021\)](#).

Figura 12 – Grafos representativos dos anos de 2007 e 2008, respectivamente - centralidade (PageRank)

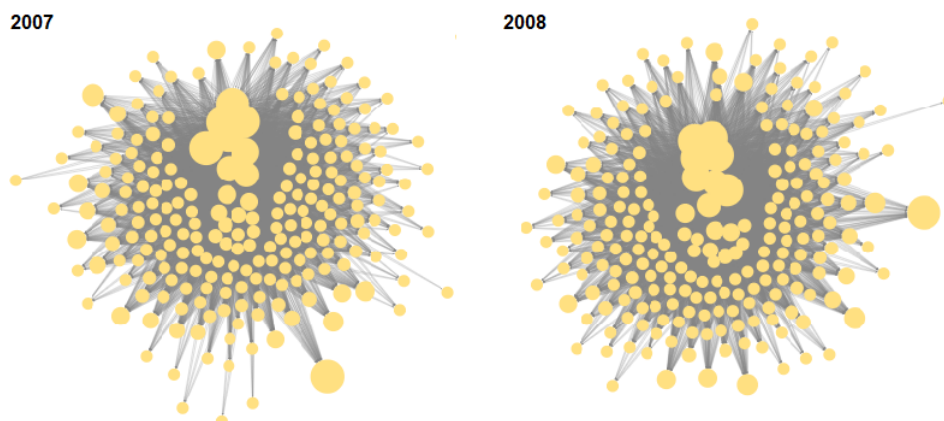
Fonte: [Desenvolvida pelo autor \(2021\)](#).

Figura 13 – Nuvem de palavras: países com maior significância de exportações nos anos de 2007 e 2008.



Fonte: UN Comtrade Database (2021).

3.2.4 Detecção de comunidades e o método Louvain

A abordagem da detecção de comunidades se dá de maneira similar à apresentada no tópico anterior. Através do Código-fonte 3 é aplicado o algoritmo de Louvain, de forma a se otimizar ao máximo a modularidade de um nó, e, definida a comunidade, esta é escrita como uma propriedade denominada `communityId` em cada um dos elementos do grafo, resultando assim na disposição apresentada na Figura 14, onde os nós são colorizados de forma a representar suas comunidades.

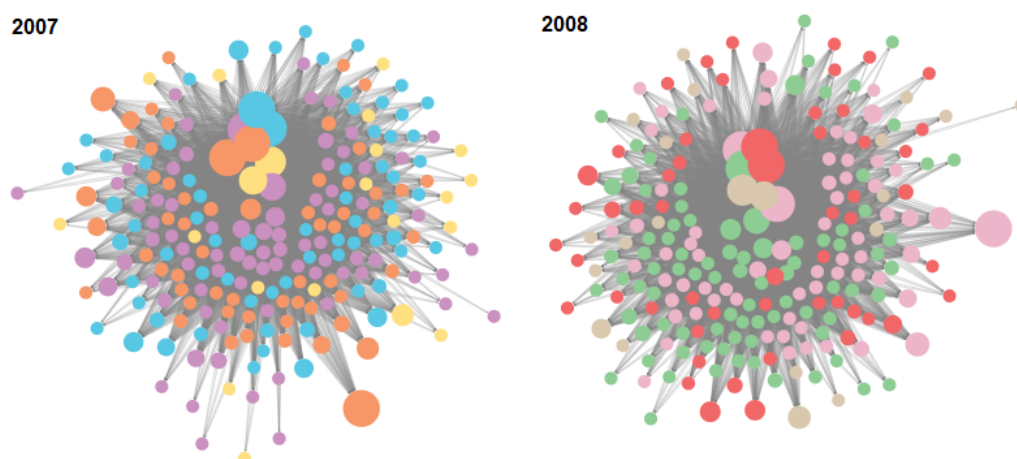
Código-fonte 3: Cypher - Louvain

```
1 CALL gds.louvain.write('Grafo_2008', {
2   relationshipWeightProperty: 'weight',
3   writeProperty: 'TestecomunityId'
4 })
```

A Figura 15 apresenta a mesma distribuição de comunidades utilizando a ferramenta gratuita Google Data Studio¹ de visualização de dados, onde é aplicada a visualização de dados por mapa.

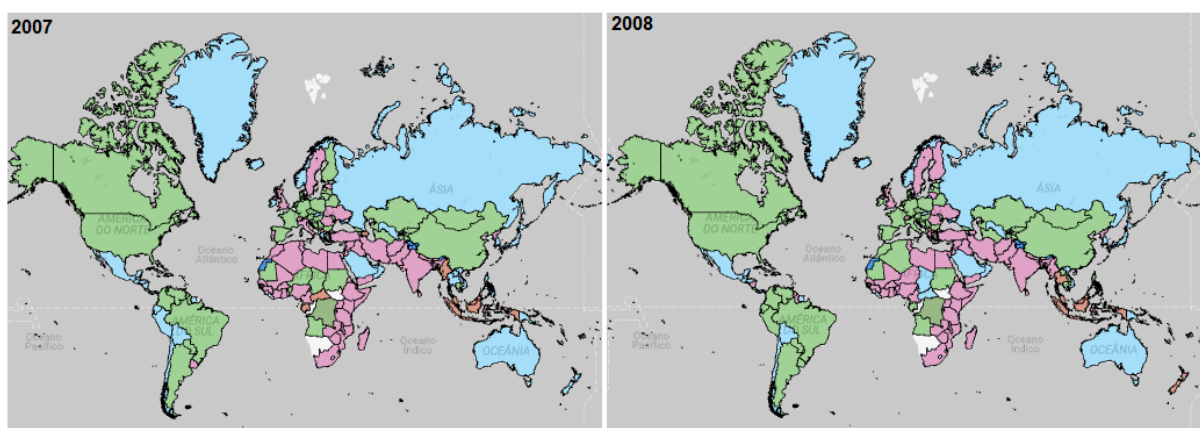
¹ <<https://datastudio.google.com/>>

Figura 14 – Grafos representativos dos anos de 2007 e 2008 - detecção de comunidades (método Louvain)



Fonte: Desenvolvida pelo autor (2021).

Figura 15 – Mapa de visualização de comunidades nos anos de 2007 e 2008

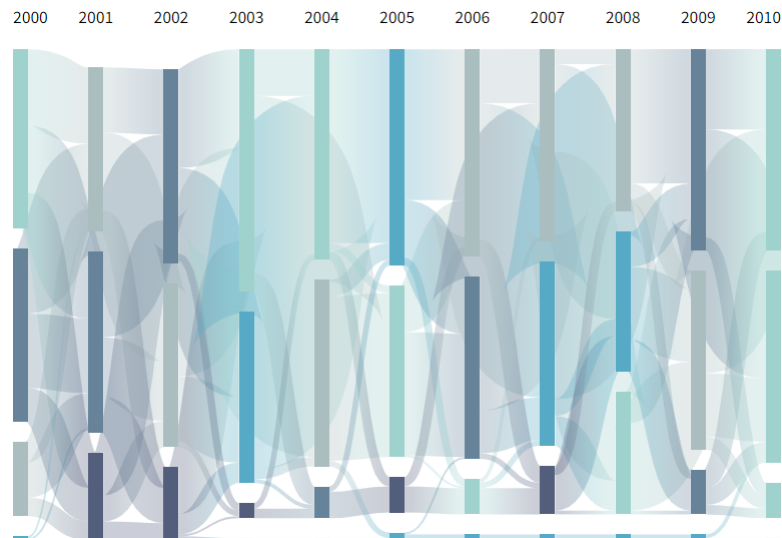


Fonte: Desenvolvida pelo autor (2021).

A detecção de comunidades por meio do algoritmo de *Louvain* resulta em distintas comunidades ao longo dos anos para os países, isso é, um agrupamento determinado em um ano pode ser diferente em um ano seguinte. Para expor essas mudanças, é apresentado o diagrama do tipo *Sankey* na Figura 16, que demonstra os fluxos de mudanças das comunidades ano a ano, as barras verticais representam as comunidades em um determinado ano e as linhas horizontais os fluxos de mudanças. A Figura 17 ressalta as mudanças ocorridas no ano de 2008 em específico, para uma das comunidades.

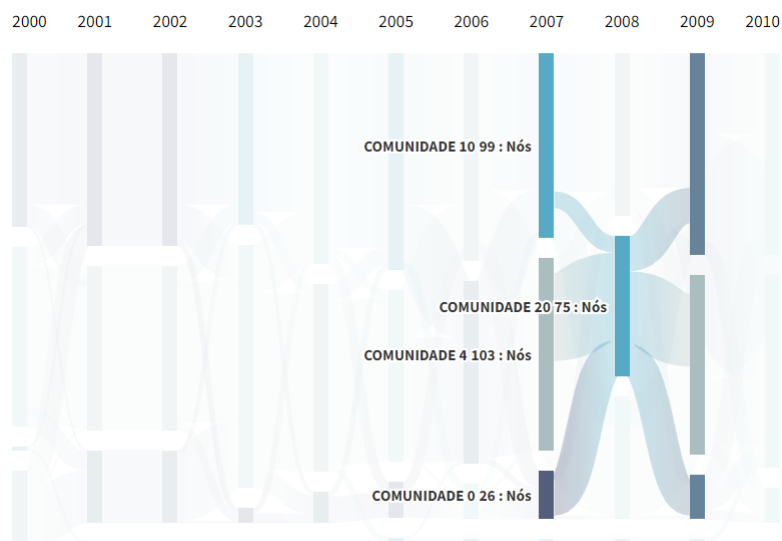
A movimentação dentre os nós nas comunidades ocorre em todos os períodos analisados. O diagrama da Figura 17 mostra essa informação, destacando as movimentações dos países que mudaram de comunidades entre os anos de 2007 e 2009, anos próximos à crise de 2008. Para uma análise um pouco mais detalhada sobre transições dos países entre comunidades no período de crise, foram observados os valores de balança comercial dos países em cada comunidade. Por meio do Código-fonte 4 foram calculados os valores das transações de exportações e importações aos níveis de país, ano e comunidade.

Figura 16 – Diagrama de Sankey - Distribuição das comunidades



Fonte: Desenvolvida pelo autor (2021).

Figura 17 – Diagrama de Sankey - Distribuição das comunidades no ano de 2008



Fonte: Desenvolvida pelo autor (2021).

Código-fonte 4: Cypher - Transações e Comunidades

```

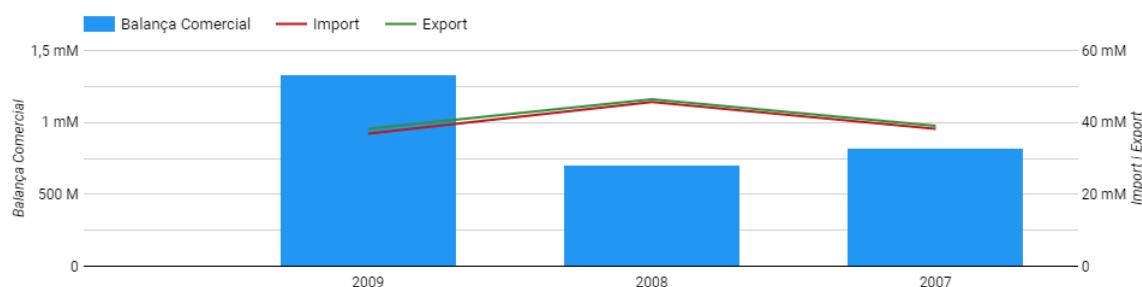
1 MATCH (n:Pais)-[r]->(m:Pais)
2 WITH n, m, SUM(r.weight) as Transações
3 RETURN n.Name, m.Name, n.Year, n.communityId ,Transações
4 ORDER BY n.Year, n.Name, m.Name

```

A partir dos dados agregados de balança comercial foram construídas as visualizações para as maiores três comunidades, dado que a quarta comunidade não apresenta movimentações nestes anos. As visualizações abordam um comparativo entre os valores médios de exportações,

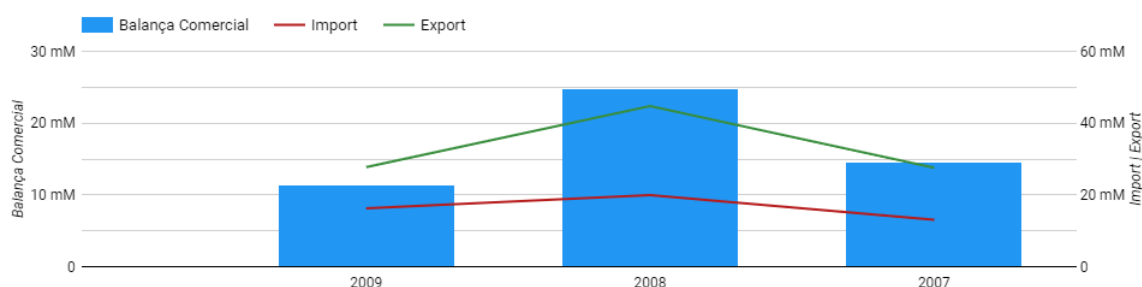
importações e balanças comerciais dos nós que permaneceram em uma mesma comunidade *versus* os que transitaram para uma nova comunidade. Essa abordagem resulta em três pares comparativos, um para cada comunidade, abaixo apresentados nas Figuras 18 e 19, respectivas à primeira comunidade. Os pares referentes às comunidade B e C são apresentados no Anexo A.

Figura 18 – Fluxos de transações dos países que se mantiveram na comunidade A



Fonte: Desenvolvida pelo autor (2021).

Figura 19 – Fluxos de transações dos países que se deslocaram na comunidade A



Fonte: Desenvolvida pelo autor (2021).

Se observa nas figuras acima que as médias das métricas dos países que transitam de comunidade se apresentam de maneiras distintas, os fluxos de importações e exportações apresentam maior distância entre suas curvas nos países que migram da comunidade em comparação aos que permanecem, configurando assim comportamentos diferentes para as respectivas balanças comerciais.

Em grande parte os países que migram de comunidade são os de pequeno porte. Para a comunidade analisada, por exemplo, Anguila, Andorra, Macedônia do Norte, Serra Leoa e Tonga se movimentam. A Figura 17 apresenta que existe um grande fluxo de mudança de comunidades, mas visivelmente a Figura 15 não apresenta este grande fluxo, e isso se dá devido ao tamanho dos países que realizam as mudanças.

Caso a caso, em sua maioria as características destes países os fazem dependentes economicamente de exportação ou turismo; A crise global de 2008 afetou fortemente ambos os setores, impactando assim na mudança da balança comercial destes pequenos países (The World Bank, 2021).

3.3 Resultados

Os resultados obtidos pelo algoritmo *PageRank* apresenta os nós referentes a países como China, Japão, Coreia do Sul, Estados Unidos, Índia e Hong Kong como sendo os com maior significância nos grafos de 2007 e 2008, com apenas alternância de posição entre Índia e Hong Kong, relação que de fato se apresenta em outras visualizações ([World Integrated Trade Solution, 2021](#)).

Já a detecção de comunidades através do método *Louvain*, apresentada de forma gráfica na Figura 15, mostra alguns países variando entre as comunidades e outros países se mantendo nas mesmas comunidades (China, Estados Unidos, etc), e é condizente com o período estudado. Essas verificações podem ser constatadas levando em consideração o fluxo de transações entre os países ([Chatham House, The Royal Institute of International Affairs, 2020](#)).

As movimentações dos nós entre as comunidades ocorrem durante todos os anos abordados nesta análise. Durante os anos próximos à crise de 2008, alguns países apresentam grande mudanças em suas balanças comerciais ([ABE, 2001](#)), fato verificado nas curvas construídas com os dados agregados dos registros de transações.

Uma avaliação geral de performance indicou que o uso das ferramentas disponibilizadas na *Neo4j Graph Platform* e da linguagem *Cypher* permitiu processamento eficiente dos dados para cálculo das métricas de interesse, levando em consideração todo o contexto e massividade conjunto de dados.

Utilizando uma configuração de máquina com processador Intel Core i5-3230M 2.60 GHz, 12 Gb de memória ram e sistema operacional Windows 10 Pro 64 bits, a operação de construção dos subgrafos contidos no grafo demandou uma média de 8 segundos de processamento para sua criação em cada um dos anos presentes na análise, fato esse que traz a facilidade para a inserção de novos blocos de análises ou novas propriedades configurando os nós presentes no grafo.

Os algoritmos para cálculo de centralidade e para o cálculo de detecção de comunidades, respectivamente *PageRank* e *Louvain*, apresentaram tempo médio de processamento de 2 segundos para cada subgrafo processado.

A base de dados processada neste projeto contém pouco mais de 64 milhões de registros, 2600 nós e 80 mil relações. O processamento desse volume de dados nos tempos apresentados condiz com a performance apresentada em ([VUKOTIC; WATT, 2014](#)).

A aplicação do Código-fonte 4 para requisição de dados e agrupamento das métricas de acordo com a granularidade solicitada gerou os resultados com tempo total de processamento de 6 mili segundos. Este valor demonstra a grande eficiência da ferramenta em determinadas requisições.

3.4 Dificuldades e Limitações

O uso da ferramenta *Neo4j* apresenta, de início, uma dificuldade, dadas as especificidades da mesma, e o uso de uma linguagem nova, como é o caso da *Cypher*, demanda tempo de estudo para entendimento do modo de operação da mesma.

A ferramenta apresenta duas situações de limitações, sendo a primeira em relação ao *bulk* de inserção de dados, que apresenta certa problemática com a quantidade de dados inseridos por vez, sendo necessário particionar o arquivo CSV, que originalmente possui pouco mais de 64 milhões de registros.

A segunda limitação é em relação ao tempo necessário para o processamento de inserção de dados. A inserção de registros na base de dados demanda tempo considerável, demandando, na análise realizada, cerca de 5 horas.

CONCLUSÃO

4.1 Contribuições

O projeto traz como sua principal contribuição o armazenamento e processamento de quantidade massiva de dados, através da ferramenta *Neo4j*, de forma ágil e eficiente. O mesmo também apresenta um método de modelagem para transformação de dados temporais em grafos temporais, abordagem que pode ser aplicada a outros conjuntos de dados temporais e problemáticas diversas.

Além disso, o projeto apresenta contribuições para o entendimento de mudanças de comportamentos de transações entre países em determinados períodos, estas movimentações de modo visual.

Por fim, o projeto trouxe para minha formação profissional grande valor, dado a grande similaridade com minha atual posição no mercado. Atualmente desenvolvo fluxos e processamentos de grandes volumes de dados em empresa do segmento energético, e o projeto contribuiu com a apresentação de uma excelente ferramenta, *Neo4j*, a qual auxiliará em meu futuro profissional, assim como o emprego de metodologia de estudo e as modelagens aplicadas.

Como trabalho futuro, se sugere a aplicação do índice HS de *commodities* como propriedades das relações entre os nós dos grafos, de modo a aprofundar a análise, permitindo maior entendimento acerca das *commodities* e o quão importantes estas são para aquela rede em determinados momentos, buscando, assim, prever movimentações comerciais e gerando engenharia de valor para determinados segmentos.

REFERÊNCIAS

ABE, M. M. A crise de 2008 e seu impacto em países economicamente dependentes de commodities. In: _____. **A Crise de 2008 e Seu Impacto em Países Economicamente Dependentes de Commodities**. [S.l.]: Fundação Getúlio Vargas Escola de Economia de São Paulo, 2001. cap. 1, p. 38. Citado na página 42.

BLONDEL, V. D.; GUILLAUME, J.-L.; LAMBIOTTE, R.; LEFEBVRE, E. **Fast unfolding of communities in large networks**. 2008. Disponível em: <<https://arxiv.org/pdf/0803.0476.pdf>>. Acesso em: 02/11/2021. Citado na página 29.

BOLDI, P.; SANTINI, M.; VIGNA, S. Pagerank as a function of the damping factor. In: _____. **PageRank as a Function of the Damping Factor**. DSI, Università degli Studi di Milano: International Conference on World Wide Web - WWW 05, 2005. cap. 1, p. 10. Citado na página 28.

BRANDÃO, L. C.; DEL-VECCHIO, R. R.; MELLO, J. C. C. B. S. de. **ESTUDOS SOBRE NOVA MEDIDA DE CENTRALIDADE EM GRAFOS: A CENTRALIDADE EM CAMADAS**. 2015. Disponível em: <<https://www.proceedings.blucher.com.br/article-details/estudos-sobre-nova-medida-de-centralidade-em-grafos-a-centralidade-em-camadas-22730>>. Acesso em: 02/11/2021. Citado na página 28.

BRESSER-PEREIRA, L. C. The 2008 financial crisis. In: _____. **The 2008 Financial Crisis**. [S.l.]: Brazilian Journal of Political Economy, 2009. cap. 1, p. 17. Citado na página 23.

BRIN, S.; PAGE, L. **The Anatomy of a Large-Scale Hypertextual Web Search Engine**. 1998. Disponível em: <<http://infolab.stanford.edu/~backrub/google.html>>. Acesso em: 02/11/2021. Citado 2 vezes nas páginas 28 e 36.

Chatham House, The Royal Institute of International Affairs. **resourcetrade.earth**. 2020. Disponível em: <<https://resourcetrade.earth/?year=2006&units=weight&autozoom=1>>. Acesso em: 05/11/2021. Citado na página 42.

Desenvolvida pelo autor. 2021. Citado 9 vezes nas páginas 34, 35, 36, 37, 39, 40, 41, 51 e 52.

ERCSEY-RAVASZ, M.; LICHTENWALTER, R. N.; CHAWLA, N. V.; TOROCZKAI, Z. **Range-limited centrality measures in complex networks**. 2012. Disponível em: <<https://www3.nd.edu/~dial/publications/ercsey2012range.pdf>>. Acesso em: 02/11/2021. Citado na página 28.

Fortune company. **Fortune 500**. 2021. Disponível em: <<https://fortune.com/fortune500/2021/search/>>. Acesso em: 02/11/2021. Citado na página 30.

GIRVAN, M.; NEWMAN, M. E. J. **Community structure in social and biological networks**. 2015. Disponível em: <<https://arxiv.org/pdf/cond-mat/0112110.pdf>>. Acesso em: 02/11/2021. Citado na página 29.

NEEDHAM, M.; HODLER, A. E. **Graph Algorithms: Practical Examples in Apache Spark Neo4j**. 1005 Gravenstein Highway North, Sebastopol, CA, United States of America: [s.n.], 2019. Disponível em: <<https://neo4j.com/blog/new-oreilly-book-graph-algorithms-spark-neo4j/>>. Acesso em: 01/11/2021. Citado 6 vezes nas páginas 25, 26, 27, 28, 29 e 30.

Neo4j company. **The World's Leading Organizations Rely on Neo4j**. 2020. Disponível em: <<https://neo4j.com/who-uses-neo4j/>>. Acesso em: 02/11/2021. Citado na página 30.

ROBINSON, I.; WEBBER, J.; EIFREM, E. Graph databases: New opportunities for connected data. In: _____. **Graph Databases: New Opportunities for Connected Data**. 1005 Gravenstein Highway North, Sebastopol, CA, United States of America: O'Reilly Media, 2015. cap. 2, p. 211. Citado na página 25.

The World Bank. **International tourism, receipts for travel items (current US\$) | Merchandise exports (current US\$) - Small states**. 2021. Disponível em: <<https://data.worldbank.org/indicator/TX.VAL.MRCH.CD.WT?locations=S1>>. Acesso em: 20/11/2021. Citado na página 41.

TOPIK, S.; MARICHAL, C.; FRANK, Z. From silver to cocaine: Latin american commodity chains and the building of the world economy, 1500-2000. In: _____. **From Silver to Cocaine: Latin American Commodity Chains and the Building of the World Economy, 1500-2000**. Durham, Carolina do Norte, EUA: Duke University Press, 2006. cap. 1, p. 377. Citado na página 23.

UN Comtrade Database. **UN Comtrade Database**. 2021. Disponível em: <<https://comtrade.un.org/>>. Acesso em: 01/11/2021. Citado 3 vezes nas páginas 33, 37 e 38.

United Nations Conference on Trade and Development. **Facts and figures on commodities and commodities trade**. 2013. Disponível em: <<https://unctad.org/press-material/facts-and-figures-commodities-and-commodities-trade>>. Acesso em: 26/10/2021. Citado na página 23.

United Nations Statistics Division. **Harmonized Commodity Description and Coding Systems (HS)**. 2018. Disponível em: <<https://unstats.un.org/unsd/tradekb/Knowledgebase/50018/Harmonized-Commodity-Description-and-Coding-Systems-HS>>. Acesso em: 26/10/2021. Citado na página 23.

_____. **Trade data processing and validation methodology**. 2019. Disponível em: <<https://unstats.un.org/unsd/tradekb/Knowledgebase/50146/Quantity-and-Weight-Data-in-UN-Comtrade>>. Acesso em: 26/10/2021. Citado na página 23.

_____. **What is UN Comtrade?** 2020. Disponível em: <<https://unstats.un.org/unsd/tradekb/Knowledgebase/50075/What-is-UN-Comtrade>>. Acesso em: 26/10/2021. Citado 2 vezes nas páginas 23 e 33.

VUKOTIC, A.; WATT, N. Neo4j in action. In: _____. **Neo4j in Action**. Shelter Island, NY, United States of America: Manning Publications, 2014. cap. 2, p. 304. Citado 2 vezes nas páginas 31 e 42.

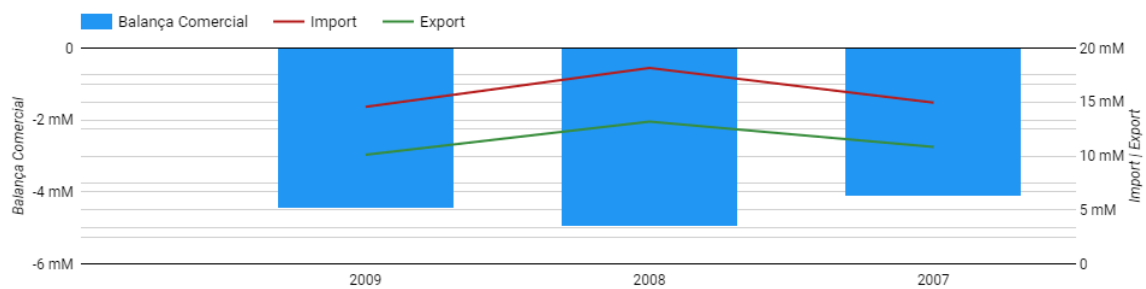
World Integrated Trade Solution . **Global perspective Total - All-Groups, for Export and Buyer 2016**. 2021. Disponível em: <<https://wits.worldbank.org/Networkchart/globalchart/en/>>

[nomenclature/H0/year/2006/tradeflow/Export/product/Total/threshold/0.95/viewpoint/Buyer>](#). Acesso em: 05/11/2021. Citado na página 42.

FLUXOS DE TRANSAÇÕES DE EXPORTAÇÕES, IMPORTAÇÕES E BALANÇAS COMERCIAIS NAS COMUNIDADES B E C

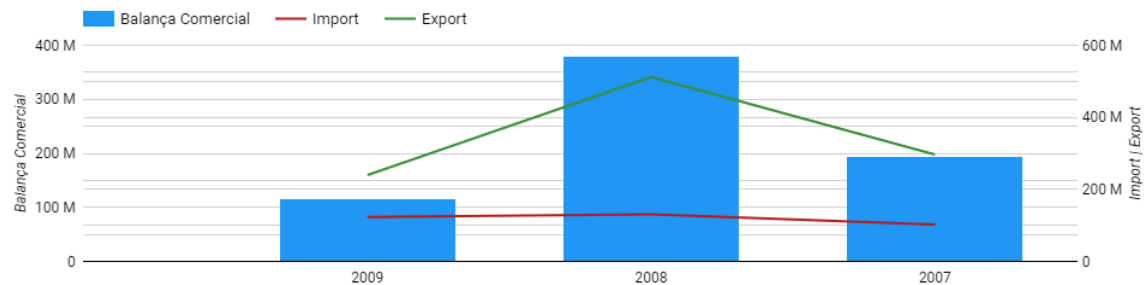
Abaixo seguem Figuras que expressam as médias dos fluxos de importações, exportações e balanças comerciais dos países que se mantiveram ou se deslocaram das comunidades B e C.

Figura 20 – Fluxos de transações dos países que se mantiveram na comunidade B



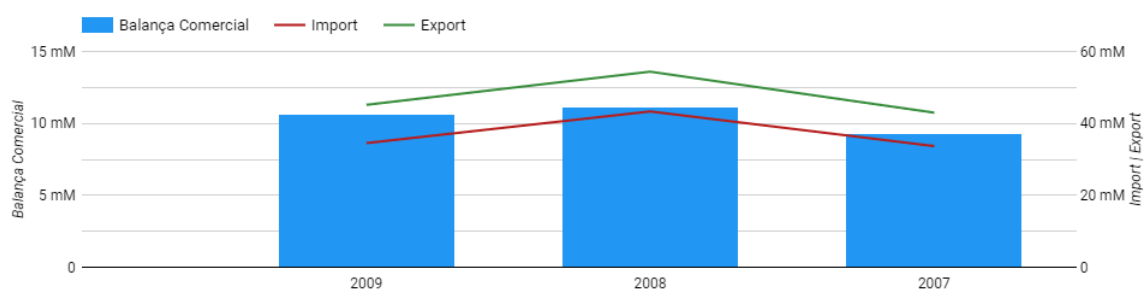
Fonte: Desenvolvida pelo autor (2021).

Figura 21 – Fluxos de transações dos países que se deslocaram na comunidade B



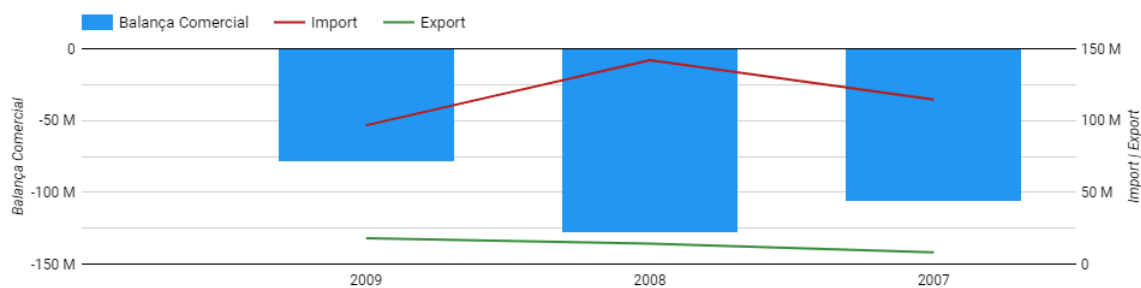
Fonte: Desenvolvida pelo autor (2021).

Figura 22 – Fluxos de transações dos países que se mantiveram na comunidade C



Fonte: Desenvolvida pelo autor (2021).

Figura 23 – Fluxos de transações dos países que se deslocaram na comunidade C



Fonte: Desenvolvida pelo autor (2021).