UNIVERSIDADE
LUSÓFONA

Análise de padrões de compra conjunta para sistemas de recomendação numa online gift store

TFC – Deisi 64

Relatório Final

Henrique Aleixo – 22103544

João Serralha - 22202133

Professora Sofia Fernandes

Trabalho Final de Curso | LEI | 2023/2024

www.ulusofona.pt

Direitos de cópia

Índice

Lista de Figuras

Lista de Tabelas

Resumo

A recomendação de produtos desempenha um papel vital no panorama do comércio eletrónico atual. À medida que as lojas online continuam a crescer, este aspeto torna-se um dos principais impulsionadores do sucesso e da competitividade no mercado. No entanto, muitos sistemas de recomendação tradicionais enfrentam desafios em termos de precisão e relevância das recomendações oferecidas. Portanto, aprimorar a precisão das recomendações tornou-se um objetivo fundamental.

O nosso objetivo foi a aplicação de análise de redes sociais com intuito de beneficiar os sistemas de recomendações numa gift store. Com este propósito, exploramos um conjunto de dados que contém mais de 500.000 registos de compras de uma loja de presentes. Após a identificação e resolução de anomalias como compras sem informação do cliente, reembolsos e a filtração de retailers, exploramos padrões de compra conjunta de produtos e usamos estes padrões para investigar semelhanças entre clientes.

Usamos os registos para formar redes, as quais explorarmos através de análise de redes sociais. Criamos uma primeira rede que relaciona produtos com outros produtos, a qual permitiu-nos explorar a interação entre produtos frequentemente comprados em conjunto. Também criamos uma rede que relaciona clientes com clientes, a qual permitiu investigar a semelhança entre clientes com base nos produtos adquiridos por estes. Utilizando a plataforma Gephi, procedemos à visualização das redes e à aplicação de métodos de deteção de comunidades que exploramos com vista à sua aplicação para sistemas de recomendação.

Abstract

The recommendation of products plays a vital role in the current landscape of e-commerce. As online stores continue to grow, this aspect becomes one of the main drivers of success and competitiveness in the market. However, many traditional recommendation systems face challenges in terms of accuracy and relevance of the recommendations offered. Therefore, improving the precision of recommendations has become a fundamental goal.

Our objective was to apply social network analysis to benefit recommendation systems in a gift store. With this purpose, we explored a dataset containing more than 500,000 purchase records from a gift shop. After identifying and resolving anomalies such as purchases without customer information, refunds, and filtering out retailers, we explored joint purchase patterns of products and used these patterns to investigate similarities among customers.

We used the purchase records to form networks, which we explored through social network analysis. We created an initial network that relates products to other products, allowing us to explore the interaction between products frequently bought together. We also created a network that relates customers to customers, enabling us to investigate the similarity between customers based on the products they purchased. Using the Gephi platform, we proceeded to visualize the networks and apply community detection methods, which we explored with a view towards their application in recommendation systems.

1 Identificação do Problema

1.1 Introdução

A recomendação de produtos desempenha um papel vital no cenário do comércio online, moldando a experiência do usuário e influenciando os resultados económicos. À medida que as lojas online continuam a proliferar, esse aspeto torna-se um dos principais impulsionadores do sucesso e da competitividade no mercado. Portanto, aprimorar a precisão das recomendações de tornou-se um objetivo fundamental.

1.2 Enquadramento Prático

Nesse contexto, a recomendação de produtos é essencial pois simplifica a experiência de compra online. As plataformas de comércio podem identificar os produtos que são mais propensos a agradar a um cliente específico com base no seu histórico de compras, preferências e comportamento de navegação. Isso não só economiza tempo para os compradores, mas também melhora sua satisfação, aumentando a probabilidade de concluir uma compra.

1.3 Descrição do Problema

Neste trabalho vamos explorar abordagens de redes sociais para sistemas de recomendação, o que envolve considerar uma variedade de algoritmos e técnicas de filtragem para analisar padrões de usuários semelhantes.

Para podermos explorar a aplicação de abordagens de redes a sistemas de recomendação, foi-nos fornecido um conjunto de dados com mais de 500.000 registos de compras de uma gift store [DataSet].

Esses registos incluem as seguintes informações:

1. InvoiceNo – número inteiro de 6 dígitos único atribuído a cada transação
2. StockCode – número inteiro de 5 dígitos único atribuído a cada produto distinto
3. Description – nome do produto
4. Quantity – quantidade de cada produto por transação
5. InvoiceDate – dia e hora da geração de cada transação
6. Unitprice – preço do produto por unidade
7. CustomerID – número inteiro de 5 dígitos único atribuído a cada cliente
8. Country – nome do país do cliente que fez a transação

Para iniciar este projeto, estabelecemos certos objetivos que pretendemos abordar durante a exploração do conjunto de dados. Começamos por identificar os problemas presentes no dataset e por contabilizar o número de produtos, clientes e compras.

Após obtermos estas métricas, iremos determinar quantas compras são habitualmente efetuadas por produto e por cliente, seguido de quantos produtos são geralmente comprados por cliente. Pretendemos também analisar o período abrangido pelos dados e observar como varia o volume de compras e de registos ao longo do mesmo.

Sumário dos Dados

Após estabelecermos os objetivos específicos para a exploração do conjunto de dados, começamos por identificar que o dataset continha 3633 produtos, 4338 clientes, 18473 compras e 540.910 registos. É importante salientar a diferença entre um registo e uma compra. Apesar de um registo representar uma compra, na realidade, uma compra pode abranger um ou vários registos. É através do 'InvoiceNo' presente em cada registo que é possível agrupar vários registos numa só compra.

Limpeza dos Dados

Na exploração e análise dos dados detetámos problemas como certos registos que não possuíam ID do cliente e outros que estavam marcados como negativos, indicando reembolsos. Estas exceções não são de relevância para o nosso projeto porque precisamos que os registos tenham ID de Cliente para formar as redes. No caso dos reembolsos, visto que não são compras realmente efetuadas, não vamos queremos utilizar para a nossa exploração.

Visto isto, optamos por remover essas instâncias. Como resultado, ficamos com 397.924 registos, os quais denominamos de dados utilizáveis. Na Tabela 1 apresentamos os valores de cada categoria de registos.

	NÚMERO DE INSTANCIAS INDIVIDUAIS
NÚMERO DE REGISTOS	541 910
REGISTOS SEM ID DO CLIENTE	135 081
REGISTOS DE REEMBOLSOS	10 625
DATA UTILIZÁVEL	397 924

Tabela 1 - Tabela dos Registos

Para melhor entender a magnitude destes valores decidimos criar o seguinte gráfico visível na Figura 1 para poder visualizar percentualmente a divisão dos registos:

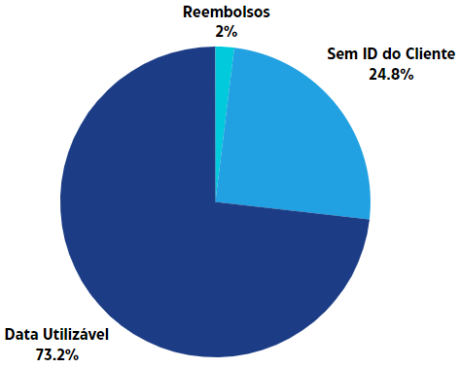

Figura 1 - Divisão dos Registos

Na Tabela 2 encontra-se os valores dos resultados obtidos, sobre o número de clientes compras e produtos, após serem removidos os dados não utilizáveis.

	DADOS FILTRADOS
NÚMERO DE CLIENTES	4338
NÚMERO DE COMPRAS	18473
NÚMERO DE PRODUTOS	3663

Tabela 2 – Sumário dos Dados

Para uma melhor compreensão dos dados, optamos por criar histogramas que nos permitam visualizar a distribuição do número de compras por produto e por cliente, assim como a distribuição de produtos por cliente e a quantidade de compras diárias. Essas representações visuais auxiliarão na identificação de padrões nos dados.

De modo a complementar os histogramas e para uma melhor interpretação decidimos, realizar o cálculo dos três quartis estatísticos, representados por Q1, Q2 e Q3. O primeiro quartil (Q1) é o valor que divide o conjunto de dados ordenados em duas partes iguais, onde 25% dos dados são menores que Q1 e 75% são maiores que Q1. O terceiro quartil (Q3) é o valor que divide os dados superiores em duas partes iguais, onde 75% dos dados são menores que Q3 e 25% são maiores. E finalmente, o segundo quartil (Q2), é a mediana dos dados, representando o ponto central do conjunto total.

Compras por Produto

Para compreender quão frequentemente os produtos eram comprados, criamos o Histograma de compras por produto (Figura 2).

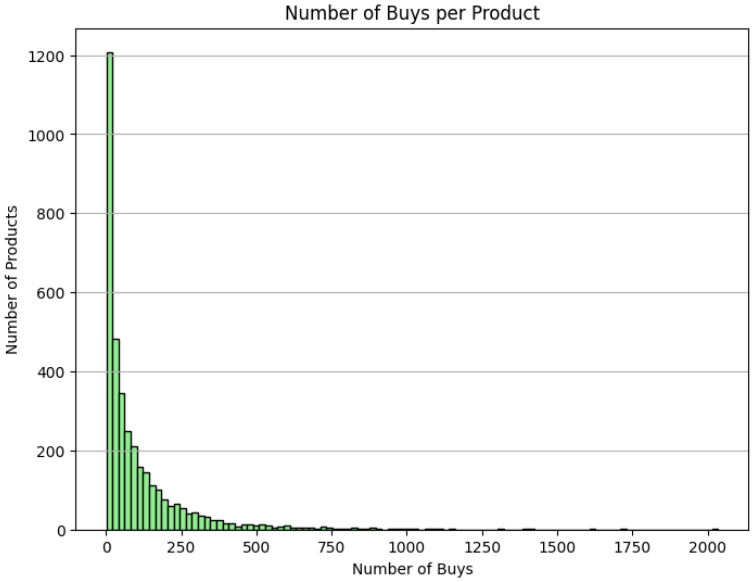

Figura 2 - Histograma Compras-Produto

Neste histograma, podemos observar que a maior parte dos produtos foram comprados poucas vezes. Sendo que o Q2 é 49, o que significa que metade dos produtos tem menos de 49 compras. Calculamos também os outros quartis tendo Q1 o valor 12 e Q3 valor 135.

É importante realçar que há alguns produtos que se destacam como 'outliers', ou seja, que foram comprados muito mais vezes que os restantes. O exemplo mais extremo é o produto 'White Hanging Heart T-Light Holder' que regista mais de 2035 compras individuais.

Compras por Cliente

Para perceber o volume de compras dos clientes criamos o Histograma compras por cliente (Figura 3).

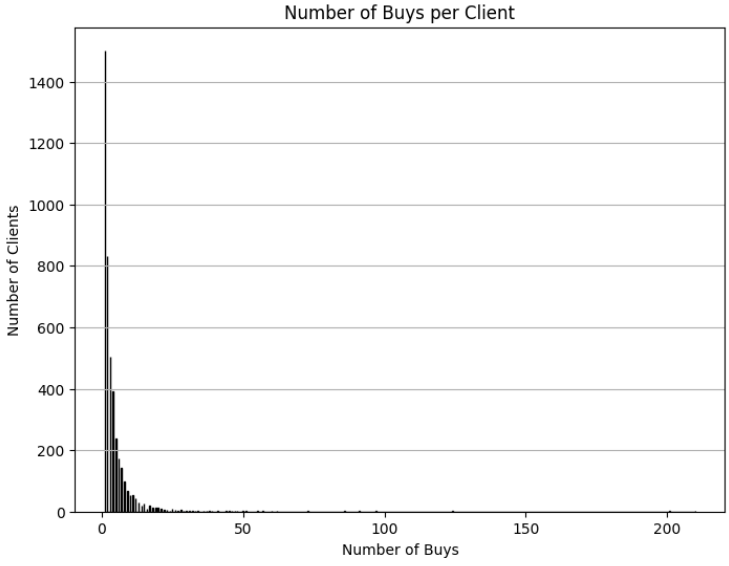

Figura 3 - Histograma Compras-Cliente

Neste histograma, é percetível que a grande maioria dos Clientes realizou poucas compras. Observamos que a mediana é 2, indicando que cinquenta por cento dos Consumidores realizaram menos de duas compras. Além disso, os outros quartis foram calculados, com Q1 a dar o valor 1 e Q3 dando 5.

Mais uma vez é importante salientar a presença de outliers, que mais adiante neste trabalho serão considerados como retailer, uma vez que o número de compras destes ultrapassa o número típico de compras de um cliente.

Produtos por Cliente

Para perceber a quantidade de produtos diferentes comprados por cliente geramos o Histograma de produtos por cliente (Figura 4).

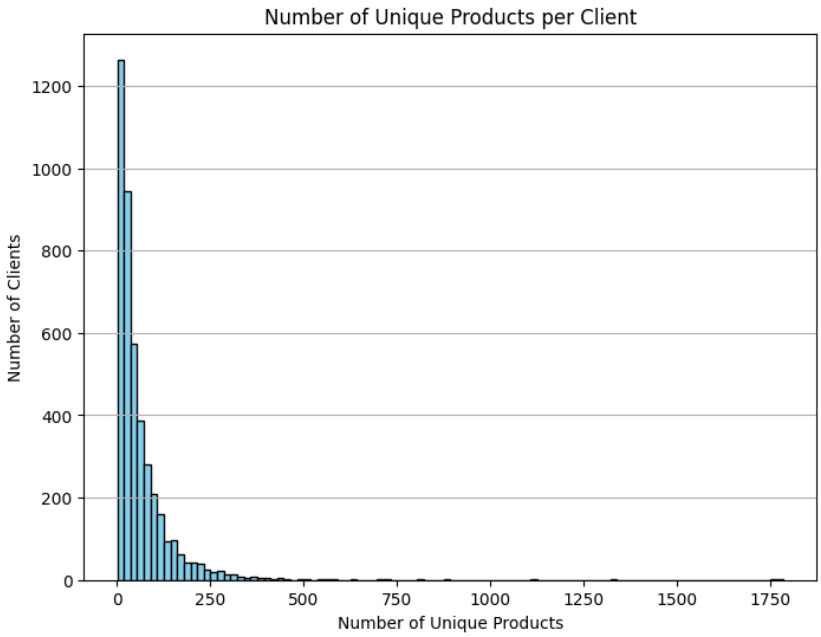

Figura 4 - Histograma Produtos-Cliente

Neste histograma, é observável que a maioria dos Clientes adquiriu uma quantidade reduzida de produtos distintos. Tendo Q2 valor 35, isso implica que metade dos Clientes comprou menos de 35 produtos distintos. Os outros quartis também foram determinados, com Q1 valor 16 e Q3 valor 77. Mais uma vez salienta-se a presença de outliers, neste caso também se pode referir que os clientes com muitos produtos diferentes comprados poderão ser retailers.

Compras Diárias

Os seguintes histogramas ilustram a variação do volume de certos indicadores ao longo do período de recolha dos dados, dando nos assim uma perspetiva diferente dos anteriores. Sendo o primeiro registo em 01/12/2010 e concluindo-se no último em 09/12/2011.

Na Figura 5 vê-se a variação do volume de compras ao longo do período total.

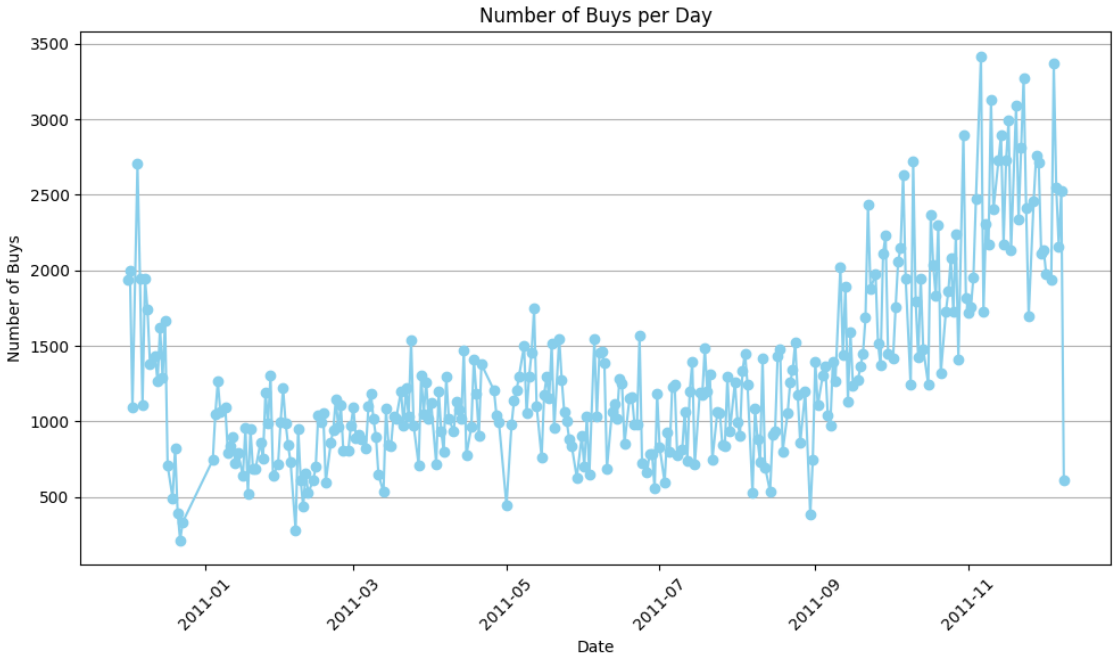

Figura 5 - Histograma Compras por Dia

Observa-se um aumento considerável no número de compras nos meses que antecedem dezembro e janeiro, o que é esperado, considerando que a loja é uma loja de presentes. Isso sugere que há um volume maior de compras durante a época festiva.

A evolução, ao longo do período dos dados, que se verifica no histograma de compras por dia é idêntica à evolução do número de diferentes produtos comprados por dia e o número de consumidores diferentes por dia.

Na Figura 6 vê-se a variação do volume de produtos comprado ao longo do período total.

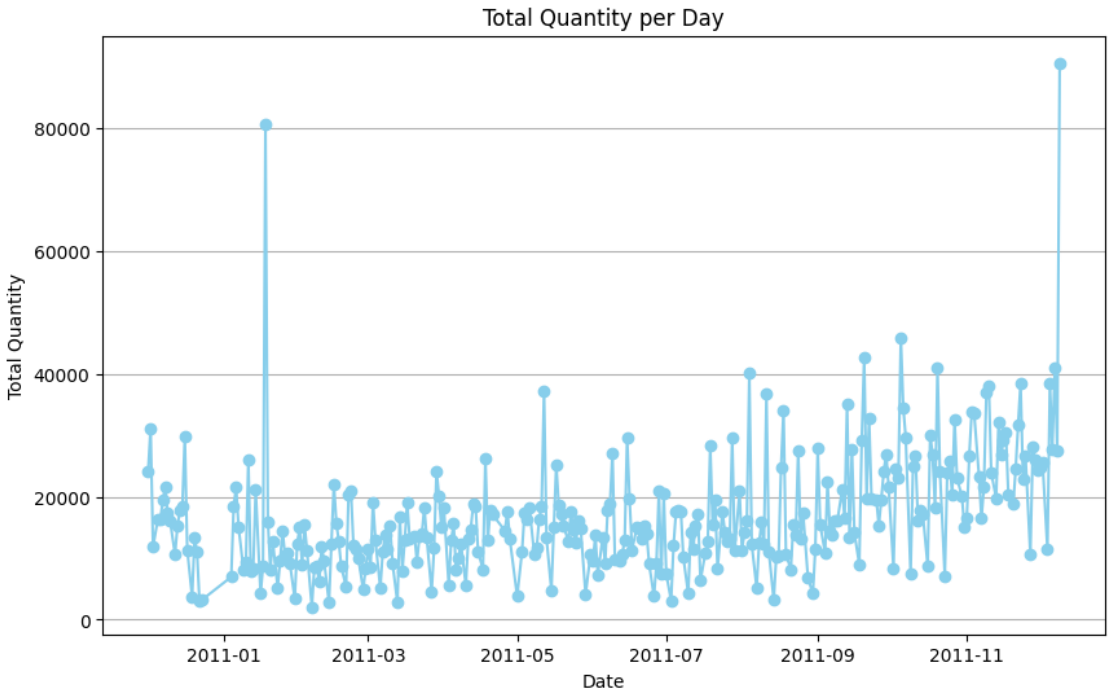

Figura 6 - Histograma Produtos por Dia

Neste histograma, é visível a presença de dias em que o volume de produtos comprados foi consideravelmente superior aos valores tipicamente observados. Esta situação é explicada pela existência dos tais 'retailers' que efetuam compras em grandes quantidades, resultando nos picos evidenciados no histograma.

Após esta análise e o processo de limpeza do conjunto de dados, tornou-se evidente a importância crítica deste fase no ciclo de vida do TFC. Ao identificar e corrigir inconsistências, outliers e dados em falta, pudemos garantir a qualidade e integridade dos dados, fornecendo assim uma base sólida para análises subsequentes. Esta componente do trabalho serviu não apenas para preparar os dados para as etapas seguintes, mas também para compreender melhor a natureza e a estrutura dos dados, destacando padrões e insights preliminares que podem beneficiar as análises exploratórias posteriores.

2 Benchmarking

2.1 Estado da Arte

Atualmente existem diferentes tipos de sistemas de recomendação [AlJa18], nomeadamente *Collaborative Filtering*, *Content-Based Filtering*, *Hybrid Systems*, *Deep learning-Based* entre outras.

Uma das abordagens principais é *Collaborative Filtering* [EkRi11], este método recomenda itens aos utilizadores com base nas interações de outros utilizadores com preferências e comportamentos semelhantes. Por outras palavras, os algoritmos de *Collaborative Filtering* agrupam utilizadores com base nos seus padrões de compra e usam as características gerais do grupo para recomendar itens a um utilizador específico. O princípio subjacente é que utilizadores com padrões de compra semelhantes tendem a ter interesses e gostos semelhantes.

A abordagem *Content-Based Filtering* [WaLi18] as recomendações baseadas em conteúdo são feitas com base nas características dos itens e nas preferências históricas do utilizador. Por exemplo, se um utilizador lê frequentemente artigos sobre tecnologia, o sistema recomendará outros artigos etiquetados com keywords relacionadas com tecnologia.

Os *Hybrid Systems* [Bu02] combinam múltiplas técnicas de recomendação para aproveitar os pontos fortes de cada abordagem, mitigando as suas fraquezas. Por exemplo, combinar filtragem colaborativa e filtragem baseada em conteúdo pode ajudar a melhorar as recomendações ao incorporar tanto as características dos itens quanto os padrões de compra dos utilizadores.

Os sistemas baseados em *Deep learning-Based* [ZhYa19] utilizam redes neurais para gerar recomendações. Estes modelos podem aprender padrões a partir de grandes conjuntos de dados, sendo eficazes na captura de relações entre utilizadores e itens.

Neste contexto, a análise de redes para sistemas de recomendação já é explorada nas abordagens de *Knowledge Graphs*, que usam teoria dos grafos e análise de redes para capturar relações e dependências entre itens e/ou utilizadores dentro da estrutura das redes.

2.2 Soluções Existentes

Dentro dos sistemas Knowledge Graph [GuZh20] existem três abordagens principais:

- Métodos Baseados em Embedding

Neste tipo de abordagem os nós (itens, utilizadores) num grafo são representados como vetores numéricos. A representação vetorial captura informações estruturais e relacionamentos entre nós, preservando propriedades importantes da rede.

- Métodos Baseados nas Conecções

Métodos baseados nas conecções exploram propriedades da rede, como medidas de centralidade ou estrutura do grafo para identificar nós ou comunidades. Ao compreender a estrutura e importância dos nós dentro da rede, esses métodos oferecem recomendações com base nos relacionamentos dos nós.

- Métodos Baseados na Propagação

Métodos baseados na propagação simulam a disseminação de informação, influência ou comportamentos por meio de um grafo. Esses métodos frequentemente envolvem modelos inspirados em processos de propagação de epidemias. Ao modelar a dinâmica da propagação de informações ou influência, estes métodos oferecem recomendações com base nos padrões previstos de propagação.

A ideia deste TFC relaciona-se com os métodos atuais porque também explora a análise de redes sociais. No entanto, o foco deste TFC é explorar como vários tipos de modelação da rede têm impacto na similaridade entre produtos e como podem ser usadas para recomendação.

3 Viabilidade e Pertinência

Ao longo deste projeto, serão exploradas metodologias para a análise das interações que se estabelecem entre utilizadores e produtos. Será foco compreender o impacto derivado da modelação dessas interações na semelhança entre utilizadores e artigos. Ambiciona-se que os resultados deste trabalho se traduzam em melhorias nos sistemas de recomendação tais como:

- Recomendações mais relevantes e personalizadas.
- Aumento de Taxa de conversão de visitantes que efetuam compras.
- Contribuição para o avanço do campo de sistemas de recomendação.

4 Solução Proposta

4.1 Introdução

O código utilizado ao longo deste projeto está disponibilizado publicamente no seguinte repositório de GitHub: https://github.com/Shadow10Z/E-Commerce [SeAl24]

Todo o código tem comentários com a sua explicação, e no repositório encontra-se o manual de instalação com todas a livrarias utilizadas e uma tabela descritiva dos registos do dataset utilizado.

4.2 Tratamento de Dados

Numa primeira etapa processamos os dados na sua totalidade agrupando a informação do dataset pelos ID dos Clientes e associando a cada cliente os seus produtos comprados (Figura 7).

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

datasetRetail = pd.read_excel('Online_retail.xlsx')

# Group the dataset by CustomerID
customers_grouped = datasetRetail.groupby('CustomerID')

# Creates an empty dictionary to store the items bought by each customer
# ID : Description, Quantity
customer_items = {}

# Iterates through each customer
for customer_id, group_data in customers_grouped:

    # Checks if customer_id is already in dictionary
    if not (customer_id in customer_items.keys()):
        customer_items[customer_id] = []

    # For each costumer_id, it adds the product to that costumer
    for i in range(len(group_data["Country"].values)):
        dataset_values = {
                          "Description": group_data["Description"].values[i],
                          "Quantity": group_data["Quantity"].values[i]
                          }

        if(group_data["Quantity"].values[i] > 0):
            customer_items[customer_id].append(dataset_values)
```

Figura 7 - Agrupamento do Dataset

Após agrupar a informação de uma maneira estruturada procedemos com a contagem dos registos totais, registos sem ID de cliente e registo reembolso (Figura 8).

```python
#---------------- Accounting total registrations, null CustomerID and Products reimbursed ----------------#

# Accouting total registrations
total_registrations = len(datasetRetail)
print(f'Total registrations: {total_registrations}')

# Checks if CustomerID is null and counts it if true
category_counts_nullIDs = (datasetRetail['CustomerID'].isna()).sum()
print(f'Number of Customers with ID null: {category_counts_nullIDs}') #135080 null CustomerID's


category_counts_products_reimbursed = (datasetRetail['Quantity'] <= 0).sum()
print(f'Number of products reimbursed: {category_counts_products_reimbursed}') #10624 products reimbursed

#---------------- Rate Products reimbursed ---------------#

reimbursement_rate = (category_counts_products_reimbursed / len(datasetRetail['InvoiceNo'])) * 100
print(f'Rate of products reimbursed: {reimbursement_rate}') #1.9604767590130447

#---------------- Rate Purchases with no CustomerID ---------------#

noClientID_rate = (category_counts_nullIDs / len(datasetRetail['CustomerID'])) * 100
print(f'Rate of purchased with no CustomerID: {noClientID_rate}') #24.926694334288598
```

```
Total registrations: 541909
Number of Customers with ID null: 135080
Number of products reimbursed: 10624
Rate of products reimbursed: 1.9604767590130447
Rate of purchased with no CustomerID: 24.926694334288598
```

Figura 8 - Contagem dos Registos

E de seguida a limpeza dos registos sem ID de Cliente, visto que não é possível utilizar os dados sem cliente associado, e de Reembolso pois não são considerados compras efetuadas (Figura 9).

```
#----------------- Filtering the Dataset -----------------#

# Filter the Dataset where 'Quantity' needs to be greater than 0 (531285) and CustomerID not null (406829).
filtered_data_quantity = datasetRetail[(datasetRetail['Quantity'] > 0) & (~datasetRetail['CustomerID'].isna())] #397924

# It also filters the Dataset where 'StockCode' can't be BANK CHARGES and POST. These refer to expenses and not products.
filtered_data_stockcode = filtered_data_quantity[~((filtered_data_quantity['StockCode'] == "BANK CHARGES") | (filtered_da

print(f'Final number of rows based on the filtration made: {len(filtered_data_stockcode)}') #396813
```

```
Final number of rows based on the filtration made: 396813
```

Figura 9 - Limpeza Registos

4.3 Criação das Redes:

4.3.1 Interação Produto-Produto

O primeiro tipo de interação que vamos explorar é a interação Produto-Produto que ocorre quando dois produtos são comprados em conjunto. A ideia é que produtos frequentemente comprados em conjunto tem à partida algum grau de similaridade, logo se um utilizador estiver a comprar um produto X, será recomendado um produto Y e Z se esses mesmos forem normalmente comprados em conjunto com o X (Figura 10).

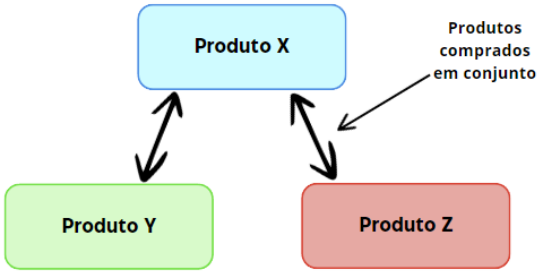

Figura 10 - Interação Produto-Produto

Com o dataset filtrado podemos agora criar as relações que vamos querer estudar. Começamos pela relação de Produto-Produto e para isso fazemos agrupamento dos dados por Cliente (CustomerID) e Produto (StockCode). De seguida vamos juntar esta tabela que criámos com ela própria sendo o ID Cliente a chave de ligação (Figura 11).

Cliente	Produto
ID_1	A
ID_1	B
ID_1	C
ID_2	A
ID_2	D

Juntar por ID Cliente +

Cliente	Produto
ID_1	A
ID_1	B
ID_1	C
ID_2	A
ID_2	D

Cliente	Produto	Produto
ID_1	A	A
ID_1	A	B
ID_1	A	C
ID_1	B	A
ID_1	B	B
ID_1	B	C
ID_1	C	A
ID_1	C	B
ID_1	C	C
ID_2	A	A
ID_2	A	D
ID_2	D	A
ID_2	D	D

Figura 11 - Merging das Tabelas

Com a relação criada podemos agora retirar as instâncias sem valor, as quais são as ligações que após o merge ligaram a compra de um produto com ele mesmo, como por exemplo 'A-A'. Ao mesmo tempo foi criada uma coluna de Peso, a qual representa o número de vezes que dois produtos foram comprados em conjunto (Figura 12).

Produto	Produto	Weight
A	B	2
A	C	2
B	C	2
A	D	2

Figura 12 - Weight

Com a tabela de relações já criadas podemos agora transformá-la num ficheiro .csv de modo a utilizá-lo no Gephi.

4.3.2 Interação Utilizador-Utilizador:

O segundo tipo de interação que vamos explorar é a semelhança entre utilizadores com base nos produtos comprados pelos mesmos. Por exemplo, temos três utilizadores, A, B e C, e verificamos que A e B compraram produtos idênticos aos do utilizador C então podemos recomendar ao usuário C produtos que ambos os outros utilizadores A e C tenham comprado, como ilustrado abaixo. Além disso, produtos comprados por A e C mas não por B, podem ser também de interesse para B (Figura 13).

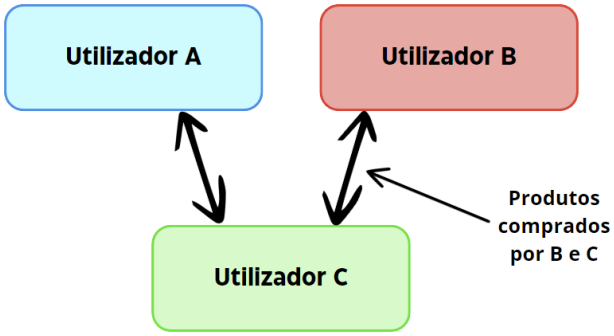

Figura 13 - Interação Utilizador-Utilizador

Para a criação da rede Cliente-Cliente o processo é análogo a da Produto-Produto sendo a única diferença que no 'Merge' utilizamos o StockCode do produto como Chave quando se junta as tabelas e depois transforma se a informação num ficheiro csv. À semelhança da rede Produto-Produto, apos obter o ficheiro, exportamos para o Gephi para prosseguir com a análise da rede.

4.3.3 Comparação e Análise das Redes

Por fim vamos analisar as várias redes criadas, e nesta análise usaremos algoritmos de deteção de comunidades [BlGu08], o que é um processo utilizado em redes para identificar grupos de nós que têm alta interconectividade entre si [Fo10]. De modo a facilitar a interpretação e análise destas comunidades faremos também manipulação do layout dos nós através de algoritmos de força. Como por exemplo *Force Atlas* [BrAn19]., *OpenOrd* [MaBr11] e *NoOverlap*. Com o mesmo intuito, vamos ajustar a espessura e opacidade da arestas e o tamanho dos nós consoante os seus pesos.

Para avaliar a qualidade da divisão das comunidades nas redes vamos usar a métrica de *modularidade* [Ne06] a qual avalia o quão bem a rede pode ser dividida em grupos de nós que estão mais fortemente conectados entre si do que com o restante da rede. Os valores de modularidade situam-se num intervalo de -1 a 1, sendo que, um valor de modularidade mais alto indica uma estrutura com comunidades bem definidas e poucas conexões entre elas, revelando uma organização interna forte.

4.4 Tecnologias e Ferramentas Utilizadas

Para análise de redes utilizamos a plataforma Gephi [BaHe09], a qual é um projeto de open Source desenvolvido para exploração e manipulação de redes.

O software Gephi têm várias funções relevantes entre as quais visualização e análise exploratória de redes, permitindo a aplicação de algoritmos de análise de redes sociais, que iremos usar. Na Figura 14 vê-se o ambiente de trabalho com um exemplo das redes criadas ao longo do projeto:.

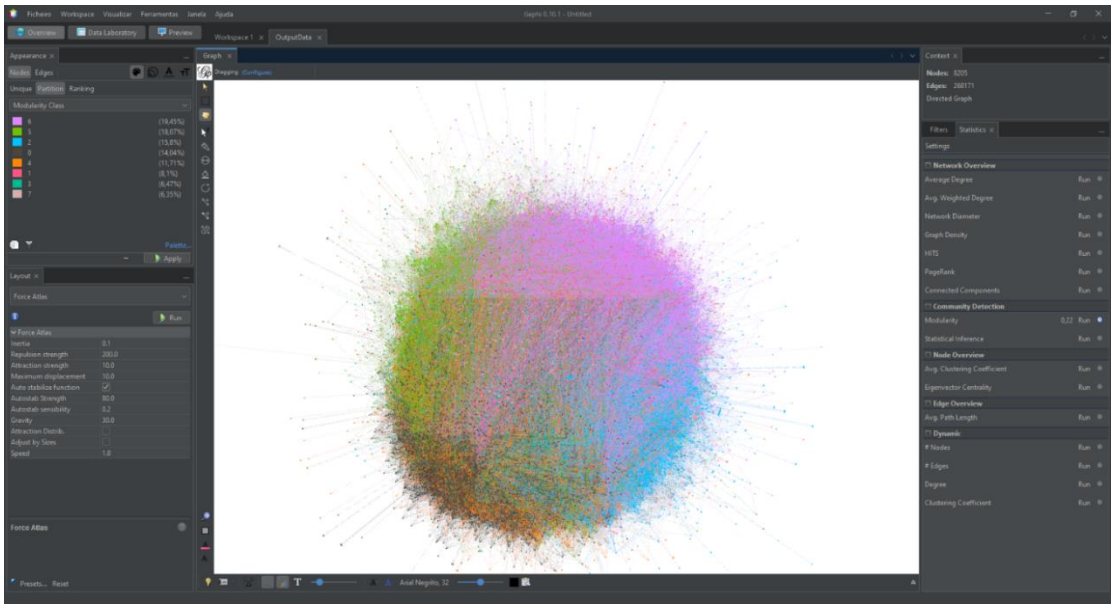

Figura 14 - Ambiente Gephi

A nível de código estamos a utilizar Python, juntamente com o Jupyter Notebook, para desenvolver todo o código necessário para tratamento de dados como queries. Importante referir que utilizamos os seguintes Packages:

- Pandas – manipulação de dados e análise dos mesmos
- Matplotlib.pyplot – biblioteca para visualização
- Seaborn – permite criação de gráficos estatísticos
- Numpy – usada para trabalhar com arrays

4.5 Abrangência

Para a execução do nosso Trabalho Final de Curso vamos utilizar várias disciplinas e áreas científicas lecionadas no curso de Engenharia Informática [DEISI24].
Sendo as mais proeminentes:

- Data Science
- Fundamentos de Programação
- Probabilidades e Estatística

Data Science foi responsável por nos ensinar a explorar dados e a familiarizar com as bibliotecas e ferramentas utilizadas ao longo do projeto.

Fundamentos de Programação proporcionou-nos com o conhecimentos base de programação o que faz de suporte para o projeto.

Probabilidades e Estatística ensinou-nos os fundamentos de análise de dados.

5 Resultados

5.1 Introdução

Com este TFC, pretendemos utilizar a informação extraída dos registos para explorar como diferentes modelações das interações podem revelar diferentes padrões nas redes criadas. Tendo em conta que queremos compreender quais as modelações mais adequadas de modo a melhorar e facilitar a análise destas mesmas redes.

Sendo este o nosso objetivo decidimos propor questões as quais queremos responder através do trabalho elaborado de modo a poder demonstrar as tais melhorias obtidas através da manipulação das redes.

5.2 Questões

Questão 1 – *Quais os produtos mais frequentemente comprados em conjunto com outros? Qual o seu valor em termos de recomendação?*

O termo ‘Product Share’ ou 'Compra em conjunto' refere-se a quantos produtos são comprados em conjunto com um dado produto. Por exemplo, se um produto A é comprado juntamente com outros 75 produtos diferentes de um total de 100, isso significa que a 'Product Share' do produto A é de 75%. Resumidamente é uma medida que mostra a popularidade de um produto.

Na Figura 15 vê-se o histograma que representa a distribuição de produtos da gift store consoante o seu valor de Compra em conjunto (na escala 0.0 é 0% e 1.0 é 100%).

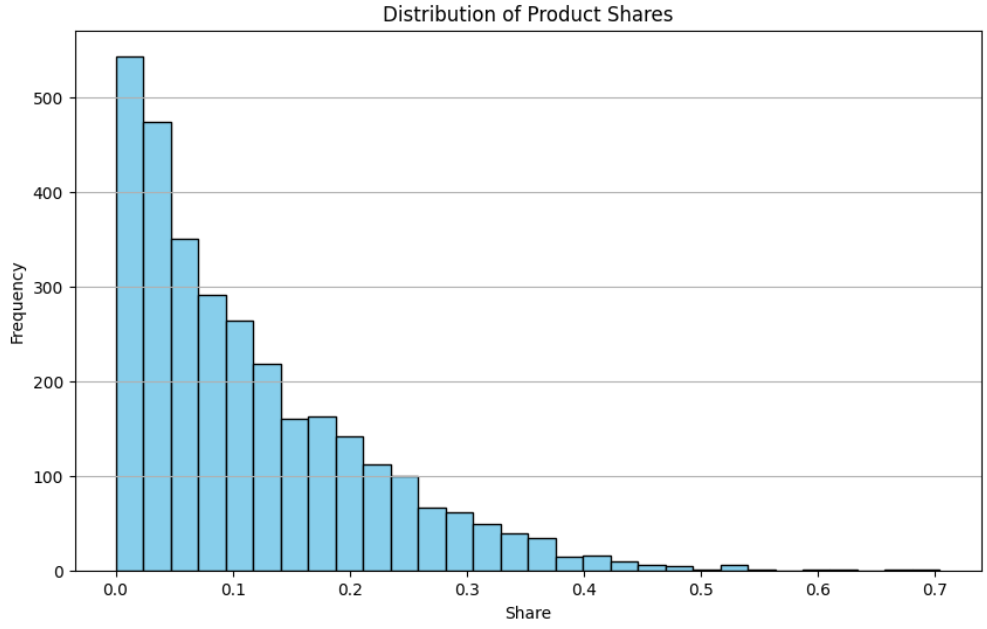

Figura 15 - Product Shares

Metade dos produtos encontram-se dentro dos valores 8% a 17%, porém como se poder observar no histograma, existem alguns casos de produtos com product share bastante alto.

Na Figura 16 vê-se apenas o número de produtos com um share superior ou igual a 40%:

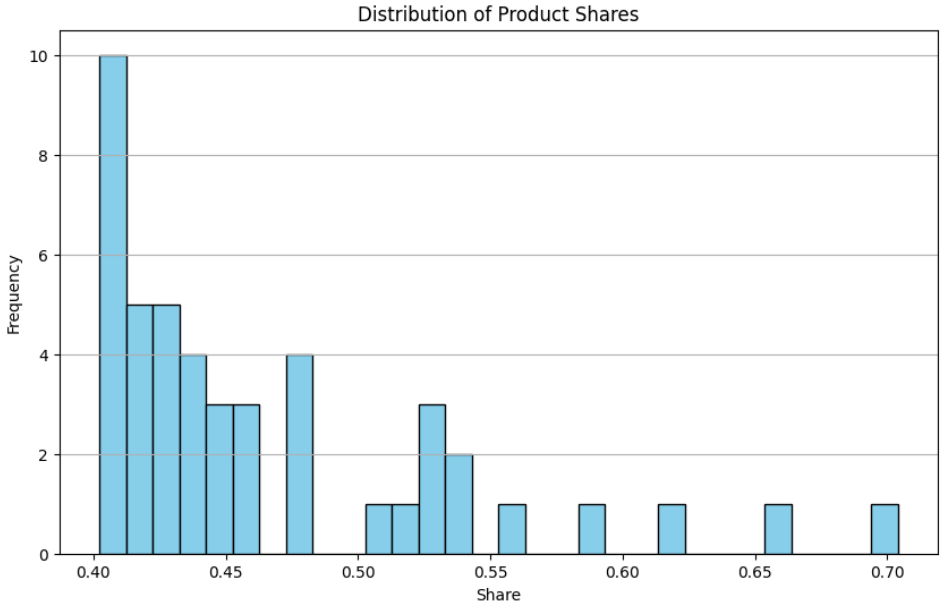

Figura 16 - Product Shares >40%

Em primeiro lugar com mais de 70% de share temos o produto 'Carry Jumbo Shopper + Rex Cash' o que é um saco de compras. É esperado que um item como um saco tenha valor elevado de product share visto que é algo que é comprado sempre independentemente dos artigos e por isso não nos ajuda a diferenciar as comunidades baseadas no género dos produtos. Em segundo e terceiro com mais de 65% de compra conjunta temos 'Hanging Heart T-Light Holder' em cores diferentes.

Estes artigos são tão frequentemente comprados com produtos de vários géneros de presentes na loja que, a nível da modelação, apresentam ligações com todas as comunidades. O que acaba por criar ligações a mais sem adicionar valor às comunidades a que estão associados. Com base na análise anterior, decidimos criar um threshold máximo de product share que um produto pode ter. Qualquer produto acima de >40% de compra conjunta será eliminado com intuito de reduzir ruído.

Na Figura 18 mostramos a rede Produto-Produto com os produtos com compra conjunta superior a 40% eliminados, sobrando 2540 nós e 96607 arestas. Cada nó, observável como um círculo, representa um produto e o tamanho dele representa a quantidade de vezes que foi comprado em conjunto com outros produtos. Estes nós estão ligados entre si através de arestas como se pode ver na Figura 17:

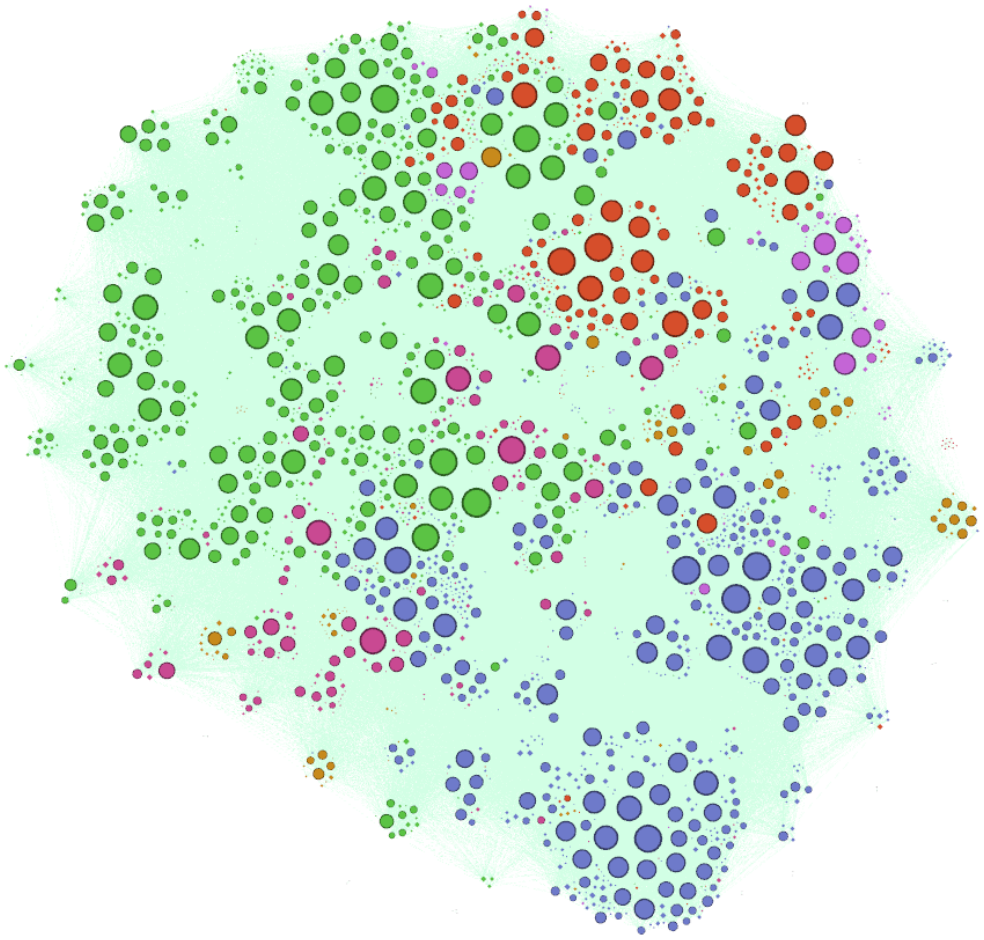

Figura 17 - Rede Produto-Produto

As arestas representam a ligação entre produtos, ou seja, dois produtos estão ligados por uma aresta se foram comprados em conjunto. A intensidade da cor e espessura da aresta é o peso da ligação, ou seja quanto mais vezes forem comprados em conjunto mais forte e visível a sua aresta de ligação.

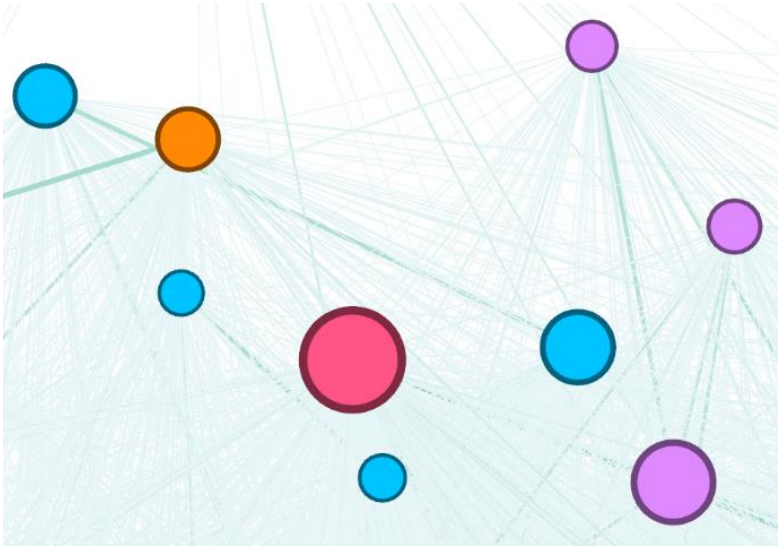

Figura 18 - Arestas Produto- Produto

Questão 2 – Quais os produtos comprados mais frequentemente em conjunto?

Na rede gerada, obtivemos um valor de modularidade de 0.23, o que é considerado um valor baixo. Esse valor indica que a estrutura de comunidade presente na rede não é muito forte, ou seja, as conexões dentro das comunidades não são consideravelmente mais densas do que as conexões entre diferentes comunidades. Apesar disso, ainda é possível visualizar algumas comunidades distintas, como é aparente na Figura 17.

Existem seis comunidades principais, sendo a distribuição de nós/produtos pelas comunidades ilustrada na Figura 19.

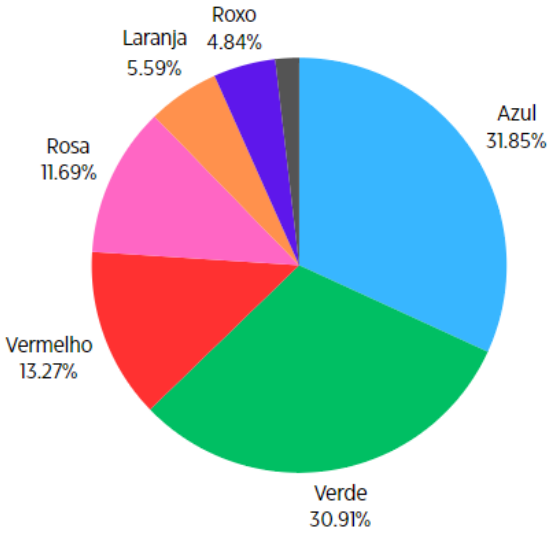

Figura 19 - Divisão Comunidades Produtos

Apesar de se destacarem estas seis comunidades principais, há 17 comunidades identificadas que não possuem peso suficiente para serem visíveis. Estas são comunidades isoladas, cada uma contendo menos de dez artigos.

Adicionalmente na Figura 20 vê-se a distribuição do número de nós por comunidade ou seja o tamanho de cada comunidade.

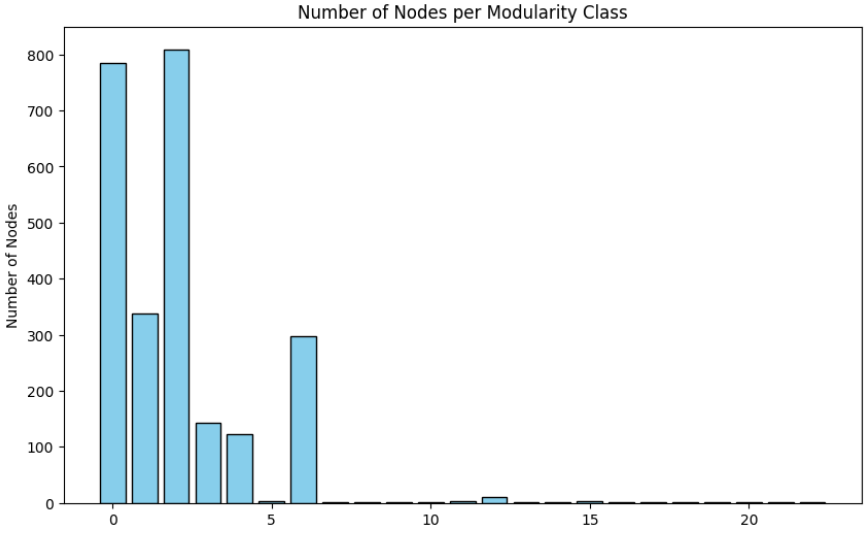

Figura 20 - Distribuição de Nós por Comunidades

Estas comunidades de produtos isolados são compostas por itens variados sendo seis destas comunidades referentes a artigos de joalheria e outras sete a artigos de decoração de casa. As restantes comunidades isoladas referem-se a produtos associados a Páscoa ou brinquedos. Estas comunidades que estamos a referir tem cada uma menos de dez elementos.

Para melhor compreender a distribuição dos nós pelas comunidades obtidas decidimos também que seria importante comparar a distribuição dos graus das várias comunidades de modo a perceber se as distribuições dos graus são semelhantes e para isso criamos os histogramas para as várias comunidades, como se pode observar na Figura 21.

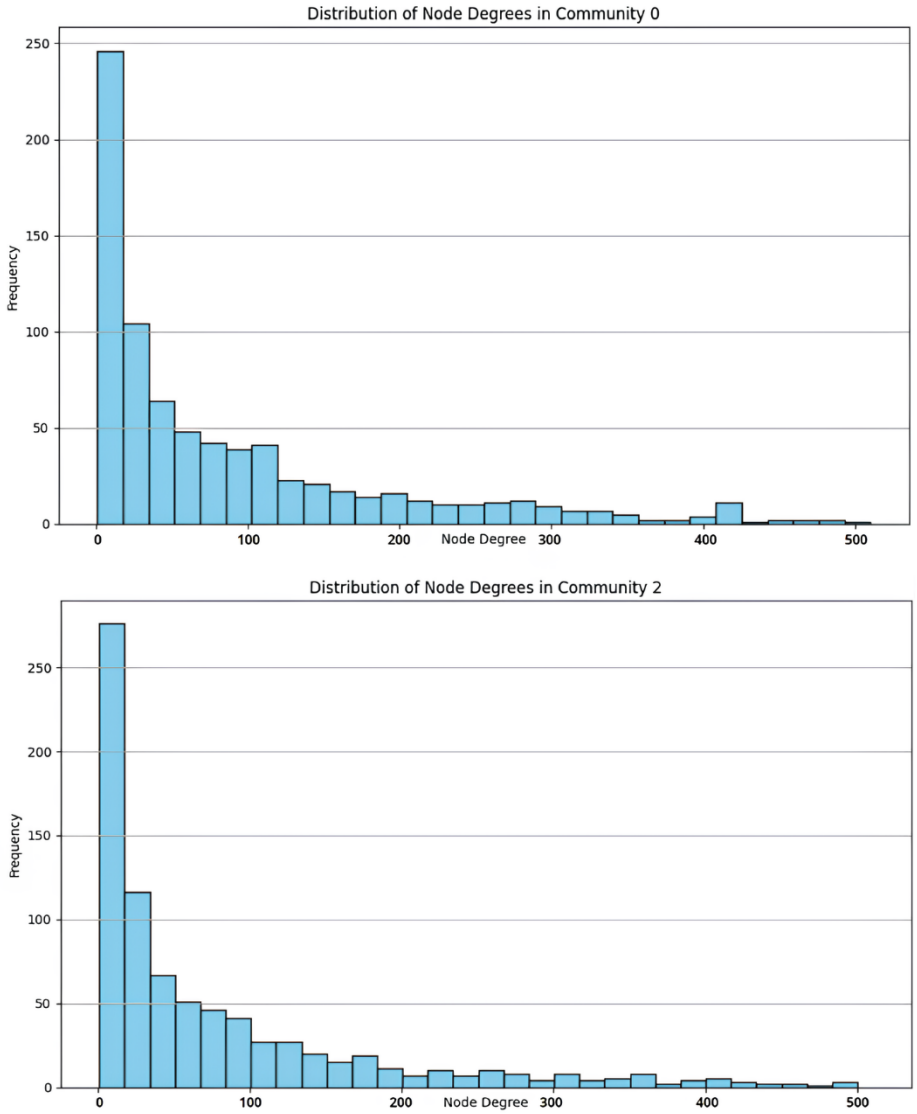

Figura 21 - Graus Comunidades Rede Produto-Produto

Como se pode observar, o padrão de distribuição dos graus dos nós nas comunidades é bastante semelhante. Observa-se um grande número de nós com um

grau baixo e uma diminuição gradual na quantidade de nós à medida que os o grau dos mesmos aumenta.

É importante notar que, para facilitar a comparação, incluímos apenas esses dois histogramas no relatório. No entanto, os histogramas do grau para as outras comunidades demonstram o mesmo padrão de distribuição podem ser encontrados no código disponibilizado.

De modo a facilitar a compreensão das diferenças entre as principais comunidades obtivemos um Top 5 dos produtos mais populares por cada comunidade, como se pode observar na Figura 22.

	Id	Description	Degree	modularity_class
0	22632	HAND WARMER RED POLKA DOT	510	0
1	21790	VINTAGE SNAP CARDS	481	0
2	22633	HAND WARMER UNION JACK	478	0
3	23356	LOVE HOT WATER BOTTLE	474	0
4	22554	PLASTERS IN TIN WOODLAND ANIMALS	465	0
5	22383	LUNCH BAG SUKI DESIGN	493	1
6	22382	LUNCH BAG SPACEBOY DESIGN	480	1
7	21985	PACK OF 12 HEARTS DESIGN TISSUES	449	1
8	22551	PLASTERS IN TIN SPACEBOY	444	1
9	20727	LUNCH BAG BLACK SKULL.	438	1
10	22624	IVORY KITCHEN SCALES	500	2
11	22776	SWEETHEART CAKESTAND 3 TIER	491	2
12	22722	SET OF 6 SPICE TINS PANTRY DESIGN	491	2
13	21181	PLEASE ONE PERSON METAL SIGN	476	2
14	21485	RETROSPOT HEART HOT WATER BOTTLE	461	2
15	22616	PACK OF 12 LONDON TISSUES	344	3
16	22998	TRAVEL CARD WALLET KEEP CALM	230	3
17	15056N	EDWARDIAN PARASOL NATURAL	224	3
18	21108	FAIRY CAKE FLANNEL ASSORTED COLOUR	211	3
19	85014B	RED RETROSPOT UMBRELLA	194	3
20	22699	ROSES REGENCY TEACUP AND SAUCER	389	4
21	22727	ALARM CLOCK BAKELIKE RED	377	4
22	22697	GREEN REGENCY TEACUP AND SAUCER	377	4
23	22726	ALARM CLOCK BAKELIKE GREEN	329	4
24	23245	SET OF 3 REGENCY CAKE TINS	315	4
25	84946	ANTIQUE SILVER TEA GLASS ETCHED	475	6
26	23322	LARGE WHITE HEART OF WICKER	457	6
27	23300	GARDENERS KNEELING PAD CUP OF TEA	440	6
28	23103	BELL HEART DECORATION	433	6
29	22578	WOODEN STAR CHRISTMAS SCANDINAVIAN	433	6

Figura 22 - Top 5 Produtos por Comunidade

Utilizando esta informação juntamente com a pesquisa manual dos produtos por comunidades conseguimos identificar os temas principais em algumas comunidades:

- **Comunidade 1 –** sacos e bolsas
- **Comunidade 2 –** utensílios de cozinha
- **Comunidade 3 –** ferramentas de gardening e outdoors
- **Comunidade 4 –** artigos de baking
- **Comunidade 6 –** objetos e decorações natalícios

Embora estes temas estejam associados aos elementos mais populares em cada comunidade existem certos produtos que não são muito semelhantes ao resto da comunidade.

Para além disso, devido ao baixo nível de modularidade inerente ao dataset, certas comunidades como a verde acabam por não ter apenas um tipo de produto em especifico mas vários temas aglomerados. Dentro da própria comunidade é possível identificarem-se várias sub-comunidades as quais têm quase que um tema próprio como se pode observar na Figura 23.

Alguns exemplos destas sub-comunidades são estes aglomerados que estão realçados, nestes conjuntos encontramos temas separados como por exemplo estes três clusters, onde o primeiro representa livros, o segundo sacos e o terceiro cartões.

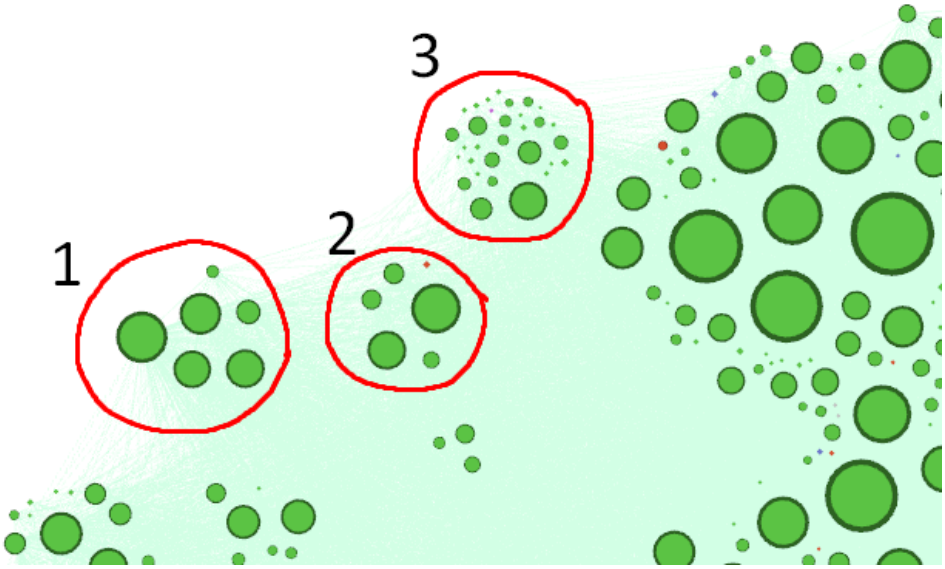

Figura 23 - Sub Comunidades

Questão 3 – *Qual o perfil dos Clientes?*

A seguinte figura representa a rede Cliente-Cliente já alterada de forma a possibilitar uma melhor visualização. Na rede cada nó é representante de um cliente e as arestas entre os nós representam a ligação entre dois clientes, ou seja clientes que tenham comprado artigos em comum. Nesta visualização é definido tamanho aos nós consoante o seu grau.

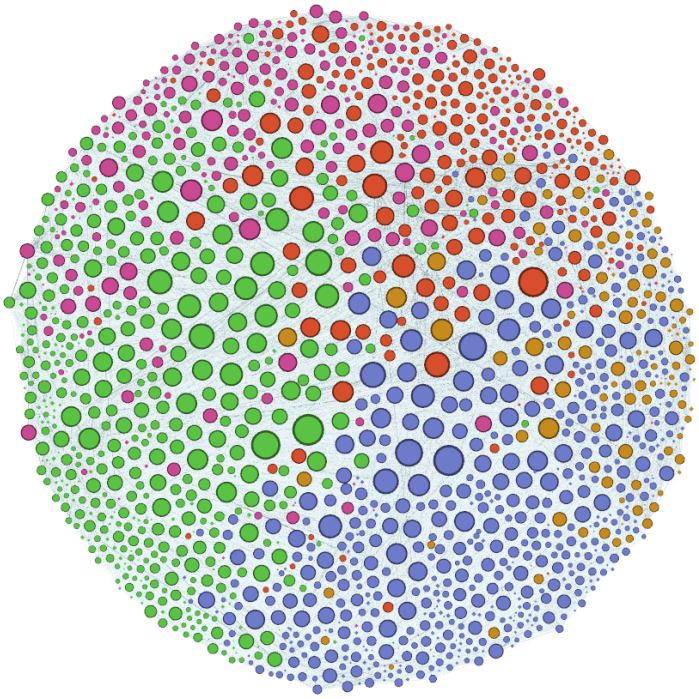

Figura 24 - Rede Cliente-Cliente

Também de modo a melhorar a visualização da rede atribuímos mais ou menos intensidade na cor e espessura das arestas entre clientes consoante o peso da ligação, ou seja clientes que tenham comprado mais produtos em comum têm uma aresta de ligação mais forte e visível como se observa na Figura 25.

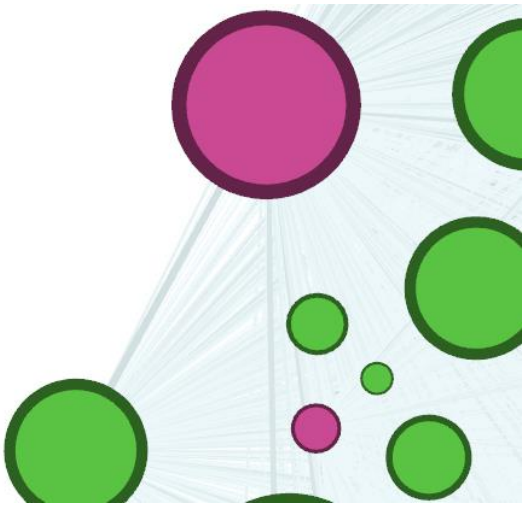

Figura 25 - Arestas da Rede Cliente-Cliente

Note se que na rede gerada obtivemos um valor de modularidade de 0.16, o que tínhamos referido anteriormente como um nível baixo de modularidade. Este baixo nível de modularidade indica que a estrutura de comunidade presente na rede não é muito forte, ou seja, as conexões dentro das comunidades não são consideravelmente mais densas do que as conexões entre diferentes comunidades. De qualquer forma ainda é possível identificar alguma divisão entre as comunidades como se observa na Figura 24.

Na rede gerada existem cinco comunidades pelas qual estão divididos os 1717 clientes. Na Figura 26 mostramos a divisão em percentagens das comunidades.

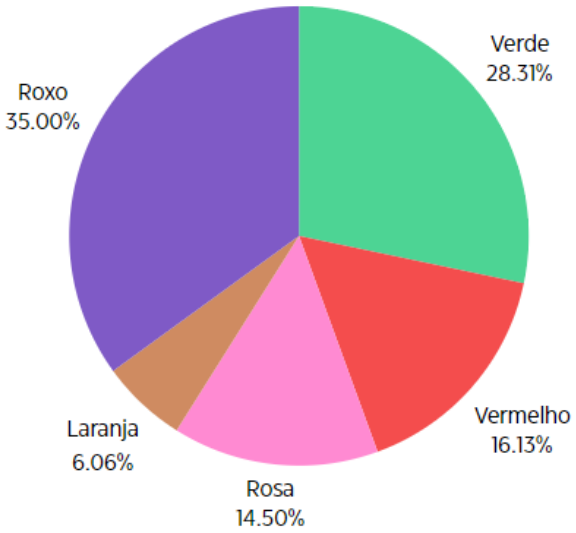

Figura 26 - Divisão Comunidades Rede Clientes

Ao contrário da rede Produto-Produto, não se verifica a presença de comunidades de clientes isoladas.

Na Figura 27 vê-se a distribuição do número de nós por comunidade, ou seja do número de clientes por comunidade.

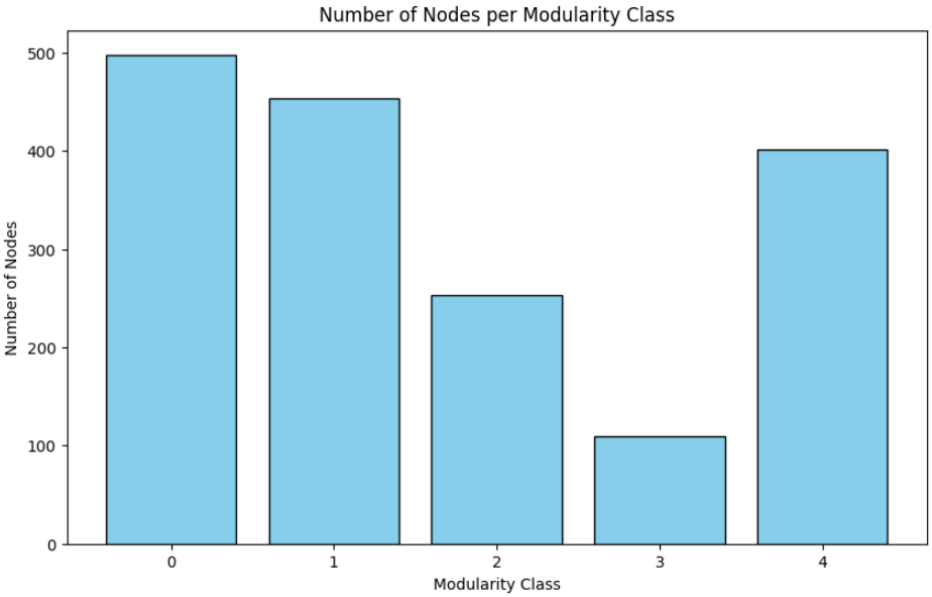

Figura 27 - Distribuição de Nós por Comunidades (Modularity Class)

No histograma da Figura 28 observa-se a distribuição total dos nós consoante o seu grau.

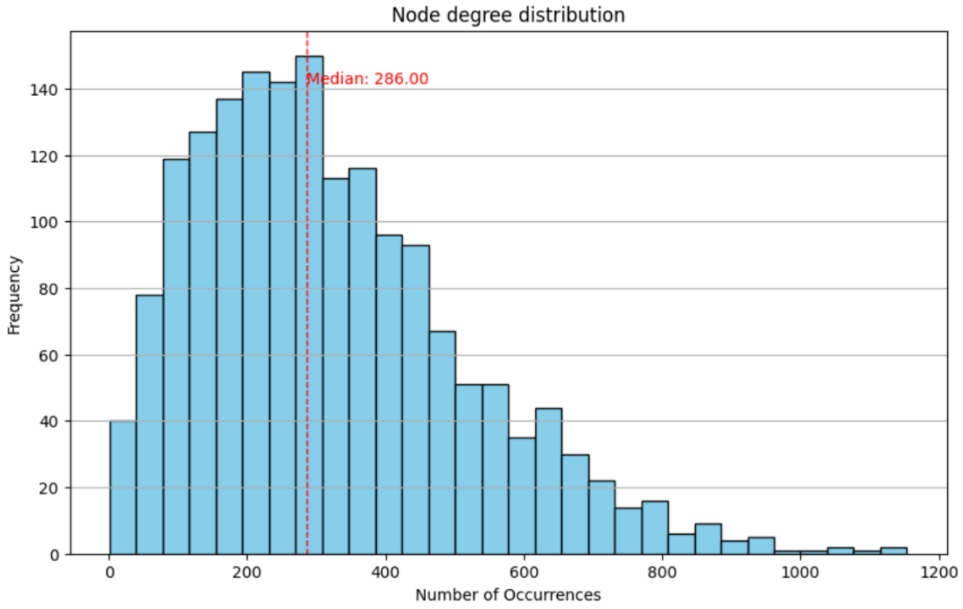

Figura 28 - Distribuição Graus dos Nós

Sendo que o grau representa o número de outros clientes que também compraram os mesmos produtos que o cliente inicial. Por outras palavras, se um cliente comprou determinados produtos, o grau do seu nó é determinado pelo número de clientes que compraram qualquer um desses mesmos produtos.

Como se observa na Figura 28, obtivemos o valor 286 como mediana. Isso significa que metade dos clientes está conectada a 286 ou mais outros clientes que compraram pelo menos um produto em comum. Essa mediana reflete a distribuição das conexões na rede de clientes, indicando que há uma grande quantidade de clientes com um número considerável de compras em comum.

À semelhança do que fizemos para a rede Produto-Produto, para melhor compreender as comunidades obtidas, comparámos a distribuição dos graus das várias comunidades da rede Cliente-Cliente de modo a perceber se as distribuições são semelhantes (Figura 29 e Figura 30).

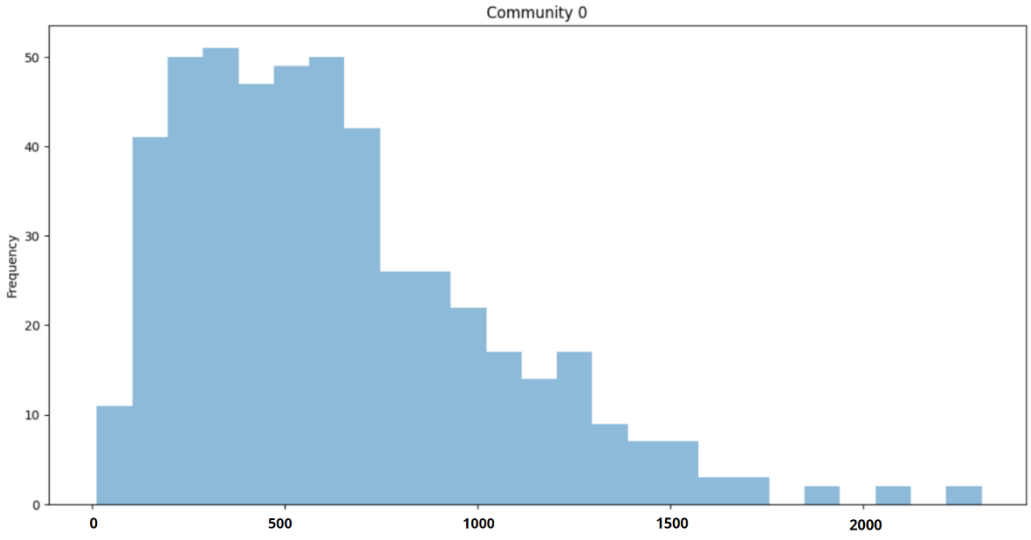

Figura 29 - Arestas Comunidade 0

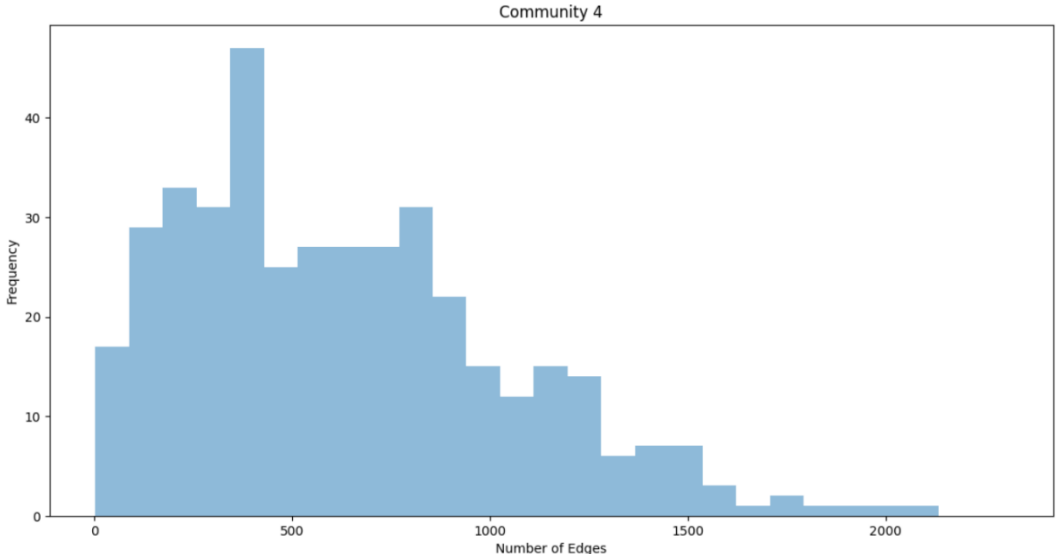

Figura 30 - Arestas Comunidade 4

Como se pode observar, o padrão de distribuição dos graus nas comunidades é bastante semelhante. Observa-se um grande número de clientes com um grau elevado e uma diminuição gradual à medida que o grau vai aumentando.

Na Figura 31 está representada a distribuição das arestas da rede Cliente-Cliente consoante o seu peso.

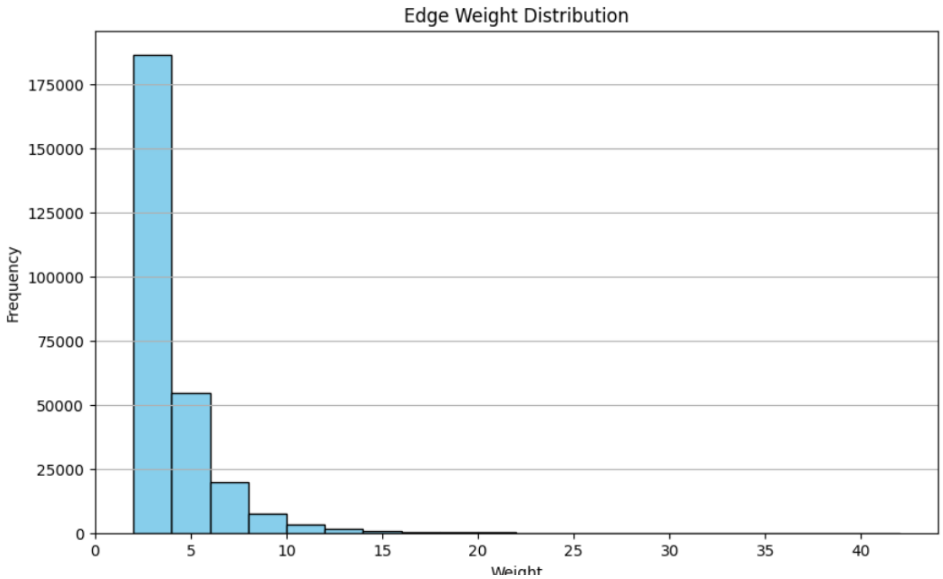

Figura 31 - Distribuição das Arestas por Peso

Como se poder ver, a vasta maioria das arestas têm um peso inferior ou igual a cinco, o que significa que uma considerável parte dos clientes que estão conectados têm menos de cinco produtos comprados em comum. Sendo que arestas com peso superior a dez são apenas uma pequena minoria. A nível de valor total de arestas da rede Cliente - Cliente temos 275 463 e um rácio de 1 nó para ≈160.4 arestas. O que é consideravelmente maior que as 96 607 arestas presentes na rede Produto-Produto e o seu respetivo rácio de 1 nó para ≈38 arestas.

6 Conclusões

Neste Trabalho Final de Curso (TFC) exploramos os registos de uma loja online de presentes, com o objetivo de observar e analisar padrões de compra conjunta. Através da análise de redes sociais pretendemos entender como vários tipos de modelação da rede têm impacto na similaridade entre produtos e clientes e como podemos usá-las para melhorar recomendações.

Com este intuito, começamos por fazer uma análise dos dados fornecidos, sendo estes os 540.910 registos de compras presentes no dataset. Descrevemos, sumarizamos e analisamos os dados. Nesta secção de análise e de limpeza deparamo-nos com problemas como clientes sem ID, reembolsos e a existência de retailers. Após a limpeza e análise dos dados ficamos com 73.5% dos registos totais e a partir destes geramos a rede Produto-Produto que explora a interação de produtos frequentemente comprados em conjunto e a rede Cliente-Cliente que explora a semelhança entre clientes com base nos produtos comprados pelos mesmos. De seguida, utilizando a plataforma Gephi visualizamos as redes e aplicamos deteção de comunidades para obtermos as suas comunidades. Analisámos os resultados através de histogramas, análise de padrões e exploração manual.

Em relação as comunidades identificadas, queremos referir o seu baixo nível de modularidade, o que indica que existe uma fraca estrutura de comunidades, o que compromete o objetivo inicial da abordagem baseada nos graus de conexão, isto porque recomendações baseadas em similaridade entre produtos ou clientes dentro de comunidades podem não ser precisas, devido as comunidades não serem claramente distintas. Este nível baixo de modularidade é mais visível na rede Cliente-Cliente a qual apenas obteve um valor de 0,16. A alta densidade de ligações entre clientes é uma das causas para a baixa modularidade e a existência de retailers poderá ser um dos motivos. Visto que retailers têm padrões de compra diferentes de um cliente individual e a quantidade/variação de compras feitas por estes retailers é bastante superior à média de um cliente individual o que aumenta o número e o peso das suas ligações, as quais já vimos que não têm valor para os nossos propósitos de recomendação.

Em suma no decorrer deste TFC, apercebemo-nos que a análise de redes sociais, embora poderosa, apresenta desafios quando a modularidade é baixa e a rede não revela comunidades bem definidas. Para além da existência dos retailers, os quais referimos anteriormente como uma das possíveis razões para um baixo nível de modularidade e criação de ruído nas redes, temos outras limitações como por exemplo a falta de informação sobre os Clientes do dataset. A única informação que temos sobre os clientes é o seu ID o que nos limita as explorações possíveis. Seria interessante ter outras informações sobre os clientes como por exemplo localização, faixa etária e preferenciais. Outra limitação é a natureza das descrições dos produtos, que são genéricas e com pouca informação limitando a exploração por NLP das descrições.

Com intuito de potenciar a análise de redes sociais para o caso desta online gift store, seria interessante explorações futuras como o refinamento da filtragem de retailers, o que pode ajudar a focar as análises nos consumidores finais, melhorando a precisão das recomendações. Outro caminho para explorações futuras seria a remoção de arestas com base no seu peso, ou seja, remover arestas com pesos mais baixos para reduzir o número de ligações de pouca importância na rede e por consequente aumentar a importância de ligações entre produtos que têm conexões mais fortes e com mais valor de recomendação.

7 Método e Planeamento

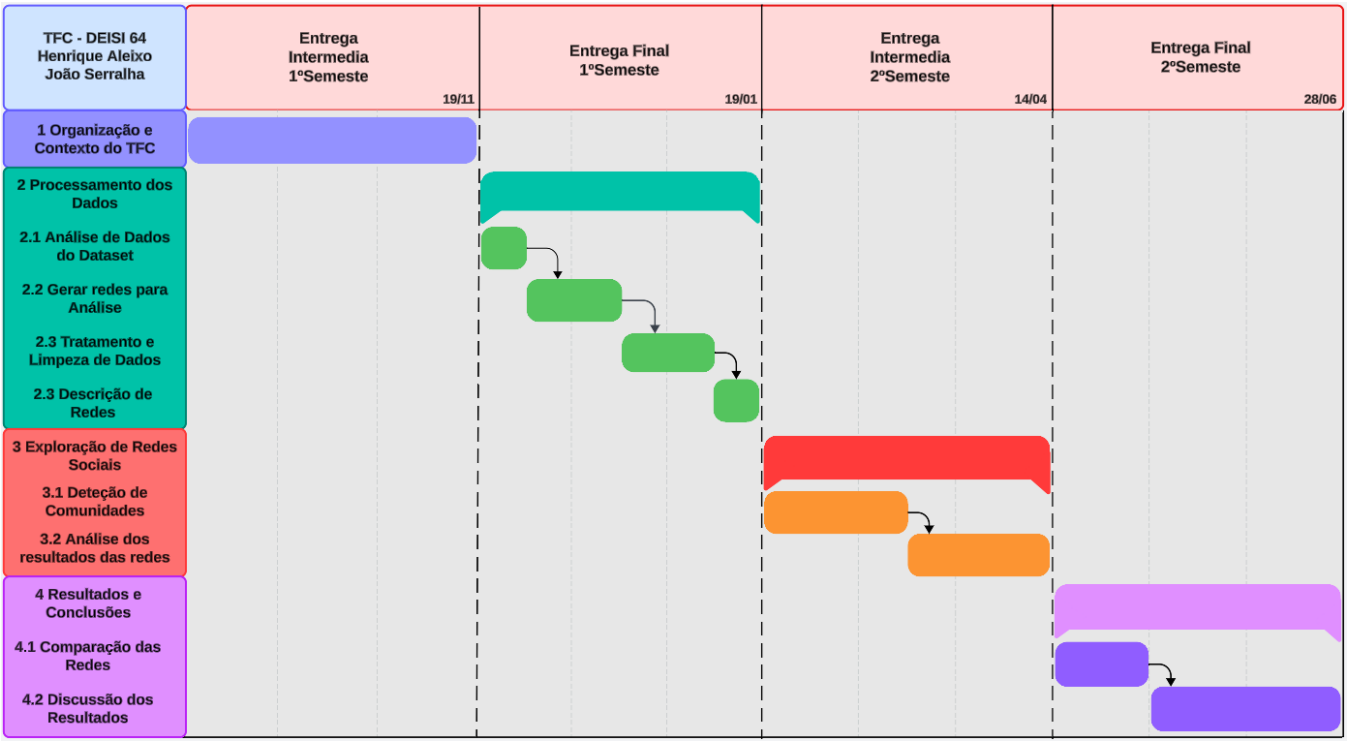

Figura 32 - Planeamento Gannt

8 Bibliografia

[DEISI24] DEISI, Regulamento de Trabalho Final de Curso, Set. 2024.

[DataSet] Online Retail. (2015). UCI Machine Learning Repository.

[GuZh20] Guo, Q., Zhuang, F., Qin, C., Zhu, H., Xie, X., Xiong, H., & He, Q. (2022). A survey on knowledge graph-based recommender systems. IEEE Transactions on Knowledge and Data Engineering, 34(8), 3549-3568.

[Fo10] Fortunato, S. (2010). Community detection in graphs. Physics reports, 486(3-5), 75-174.

[AlJa18] Alyari, F., & Jafari Navimipour, N. (2018). Recommender systems: a systematic review of the state of the art literature and suggestions for future research. Kybernetes, 47(5), 985-1017.

[Gr24] Grandjean, M. (2024). Gephi Communities Visualisation [Image]. MartinGrandjean. https://www.martingrandjean.ch/introduction-to-network-visualization-gephi/

[SeAl24] Serralha, J. e Aleixo, H. (2024). Project Repository: A collection of code and data for the research. GitHub. https://github.com/Shadow10Z/E-Commerce

[BaHe09] Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

[BlGu08] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10), P10008.

[BrAn19] Broasca, L., Ancusa, V. M., & Ciocarlie, H. (2019, October). A qualitative analysis on force directed network visualization tools in the context of large complex networks. In 2019 23rd International Conference on System Theory, Control and Computing (ICSTCC) (pp. 656-661). IEEE.

[EkRi11] Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). Collaborative filtering recommender systems. Foundations and Trends® in Human–Computer Interaction, 4(2), 81-173.

[WaLi18] Wang, D., Liang, Y., Xu, D., Feng, X., & Guan, R. (2018). A content-based recommender system for computer science publications. Knowledge-Based Systems, 157, 1-9.

[Bu02] Burke, R. (2002). Hybrid recommender systems: Survey and experiments. User modeling and user-adapted interaction, 12, 331-370.

[ZhYa19] Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. ACM computing surveys (CSUR), 52(1), 1-38.

[Ne06] Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, *103*(23), 8577-8582.

[MaBr11] Martin, S., Brown, W. M., Klavans, R., & Boyack, K. W. (2011, January). OpenOrd: an open-source toolbox for large graph layout. In *Visualization and data analysis 2011* (Vol. 7868, pp. 45-55). SPIE.

9 Glossário

LEI	Licenciatura em Engenharia Informática
LIG	Licenciatura em Informática de Gestão
TFC	Trabalho Final de Curso
KG	Knowledge Graphs
NLP	Natural Language Processing
Q1	Quartile One
ID	Identification