UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Caracterização da estrutura do comércio eletrônico

Flávio Heleno Batista

Caracterização da estrutura do comércio eletrônico

Flávio Heleno Batista

Orientador: Prof. Dr. Francisco Aparecido Rodrigues

Monografia de conclusão de curso apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP - para obtenção do título de Bacharel em Ciências de Computação.

Área de Concentração: Redes complexas

USP – São Carlos
Outubro de 2012

"A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê."

Arthur Schopenhauer

Dedicatória

Dedico este trabalho a todos aqueles que, em todos os momentos, estiveram presentes em minha vida, nos bons e maus momentos. São eles que me dão forças para continuar.

Agradecimentos

Agradeço a todo o apoio, esforço, dedicação e paciência desprendida pelo meu orientador, o Prof. Dr. Francisco, sem o qual este trabalho não existiria. Agradeço também à minha família e aos meus amigos, por me apoiarem em todos os momentos.

Resumo

Este trabalho utiliza o site de e-commerce da Amazon para gerar uma rede complexa, que representa a relação entre os produtos que são frequentemente comprados em conjunto pelos consumidores. Durante o desenvolvimento deste projeto, serão empregadas algumas medidas simples para caracterização topológica de redes complexas. A análise e caracterização da rede formada pode levar a um melhor entendimento sobre o comportamento de compra, permitindo melhorar, por exemplo, a oferta de produtos de modo que a taxa de conversão de compras aumente. Por fim, com a conclusão deste trabalho, espera-se que os resultados obtidos permitam entender melhor o comportamento do consumidor em sites de comércio eletrônico.

Sumário

Lista de Gráficos

Lista de Tabelas

Lista de Figuras

CAPÍTULO 1: INTRODUÇÃO

1.1. Contextualização e Motivação

No primeiro semestre de 2012, o comércio eletrônico brasileiro teve um faturamento de mais de R$10 bilhões, de acordo com o 26° relatório WebShoppers realizado pela empresa e-bit, com apoio da Câmara Brasileira de Comércio Eletrônico[1]. Este total representa pouco mais de 50% do faturamento de todo o ano de 2011, fato que comprova o crescimento do setor.

O sites de e-commerce se tornam cada vez mais populares e se consolidam como opção de compra para o brasileiro. Assim como em supermercados e lojas de departamento, é necessário oferecer produtos de maneira que leve o consumidor a comprar mais itens além do produto que o levou inicialmente ao site, sendo necessário entender o comportamento de compra do usuário para fazer ofertas que tenham maior probabilidade de serem transformadas em compras.

Para compreender melhor o comportamento de compra, é necessário usar alguma ferramenta que permita representar a relação entre os produtos frequentemente comprados em conjunto e os produtos propriamente ditos. Uma ferramenta que permite essa representação e o estudo das relações presentes nessa representação é definida pela teoria das redes complexas.

A teoria das redes complexas é uma área da ciência que permite generalizar a representação dos mais diversos sistemas complexos, com ferramentas universais e conceitos bem definidos na análise desses sistemas. Com aplicações que vão desde a neurociência até a sociologia [1], essa teoria apresenta caráter altamente multidisciplinar. Exemplos de redes complexas abragem as conexões entre neurônios, as interações celulares e a Internet [1]. Na figura 1 são apresentados dois exemplos de sistemas complexos, a Internet e as conexões entre neurônios.

1 http://www.e-commerce.org.br/stats.php acessado em 01 de novembro de 2012.

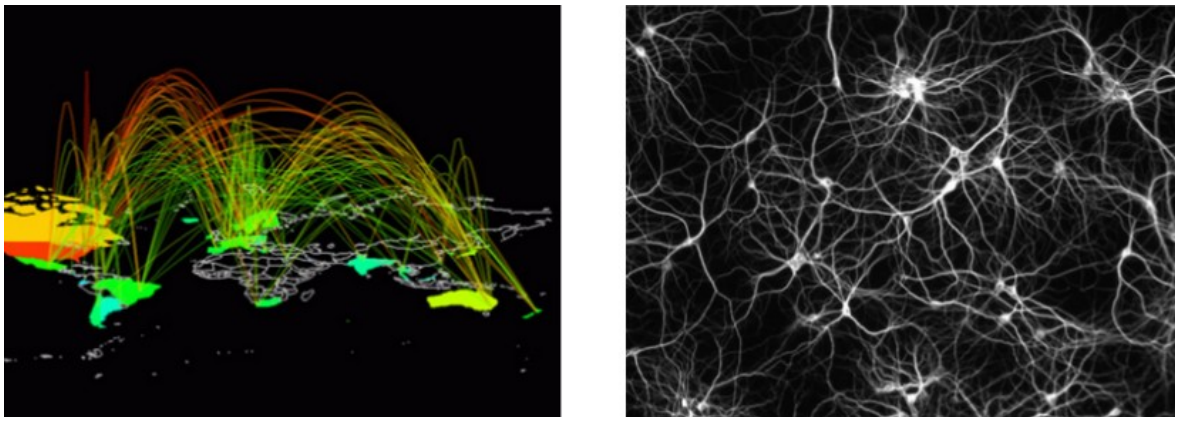


Figura 1: A internet (à esquerda), e o cérebro (à direita), constituem exemplos de redes complexas. Figuras extraídas de http://www.visualcomplexity.com.

Os sistemas complexos podem ser representado por um conjunto de elementos que se conectam de acordo com alguma interação específica, ou seja, por redes (grafos com estrutura irregular). Apesar da crença que tais sistemas possuíam uma organização aleatória, coletas de dados mostraram que a organização das conexões desses sistemas está fortemente ligada à sua natureza e às leis específicas de sua evolução [2, 3, 4]. Um exemplo prático pode ser observado nas redes sociais como Facebook e Myspace [5], onde há certas pessoas com um número elevado de amigos, enquanto a maioria das pessoas se organiza em pequenos grupos. De maneira semelhante, no caso da Internet, os roteadores localizados em grandes centros urbanos e por consequência mais importantes, possuem mais conexões do que os demais roteadores e ao longo do tempo, apresentam uma tendência a receber um maior número de conexões do que os roteadores menos conectados [4].

Por fim, adota-se uma ferramenta que permite representar a relação interesse deste trabalho e a partir dessa representação, pode-se caracterizar o comportamento da rede e a partir dessas informações obter uma melhor compreensão sobre o comportamento de compra dos usuários.

No primeiro projeto supervisionado do aluno, foi instanciado um processo completo de Mineração de Textos com base nos dados provenientes da rede social Twitter, compreendendo as etapas de identificação do problema, pré-processamento, extração de

padrões, pós-processamento e utilização do conhecimento. Como resultado foi criada uma ferramenta WEB, ainda não disponível publicamente, para a análise exploratória dos textos. Este trabalho não é uma continuação do primeiro trabalho.

1.2. Objetivos

Esse projeto em como objetivo representar um site de e-commerce através de uma rede complexa com base na relação de produtos frequentemente comprados em conjunto.

Este trabalho também tem como objetivo caracterizar a rede complexa formada com base nas medidas adotadas.

1.3. Organização do Trabalho

Esse trabalho está organizado em 4 capítulos. No segundo capítulo é feita uma revisão bibliográfica do assunto tratado: as Redes Complexas, resumindo seus conceitos básicos, representação e medidas de caracterização utilizados neste projeto. No terceiro capítulo são descritas as atividades realizadas para a conclusão do trabalho. No último capítulo são feitas as conclusões sobre o projeto e sobre a graduação do aluno.

CAPÍTULO 2: REDES COMPLEXAS

2.1. Considerações Iniciais

As relações presentes nos sites e-commerce também podem ser representadas através de uma rede complexa. Nesse trabalho foi representada a relação entre os produtos que frequentemente são comprados em conjunto, e a partir dessa representação foram empregadas algumas medidas para a caracterização de tal rede.

2.2. Conceitos básicos

Um grafo $G = (N, E)$, é formado por um conjunto de nós ou vértices ($N = \{n_1, n_2, \ldots, n_n\}$) e por um conjunto de arestas ($E = \{e_1, e_2, \ldots, e_m\}$), além disso as arestas ainda podem ser direcionadas ou não-direcionadas. Além dessas informações, a rede pode apresentar um valor de intensidade para as arestas, também conhecido como peso, e neste caso, o grafo possui ainda o conjunto $W = \{w_1, w_2, \ldots, w_m\}$, sendo representado por $G = (N, E, W)$.

Na teoria das redes complexas, sistemas complexos são representados por meio de grafos devido a algum tipo de interação [3]. Essas redes podem ser estáticas, quando não há mudança no número de nós, arestas ou mesmo mudanças na configuração das ligações entre os nós; ou dinâmicas, neste caso sendo possível modelar o seu crescimento pela análise da mudança de sua estrutura no tempo. Apesar das redes reais serem dinâmicas, elas podem ser analisadas como redes estáticas dentro de um intervalo de tempo em que as mudanças não são importantes ou não ocorrem.

2.3. Representação

Computacionalmente, os grafos, e por consequência as redes complexas, podem ter sua estrutura representada através de uma matriz de adjacência ou por uma lista de conexões. Na lista de conexões, apenas os pares de vértices (i, j) que possuem ligações são armazenados. Na matriz de adjacência, se dois vértices i e j estão ligados entre si, adota-se o valor 1 para a entrada a_{ij} na matriz ou o valor 0, caso contrário. A figura 2 demonstra a

representação de uma rede não direcionada e de uma rede direcionada, através de uma matriz de adjacência.

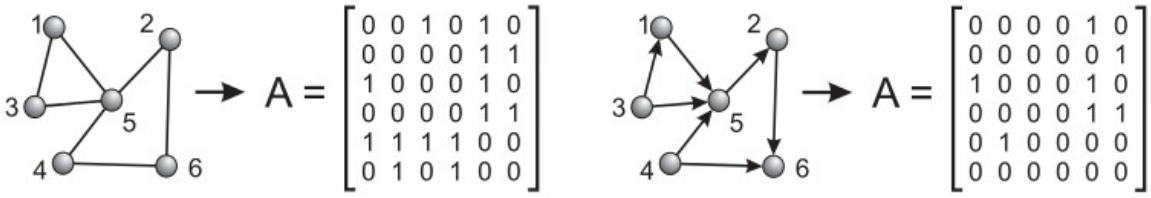


Figura 2: Exemplo de uma rede não direcionada (à esquerda) e uma rede direcionada (à direita). No primeiro caso o elemento a_{ij} da Matriz A é 1 se existe ligação entre os vértices i e j e 0 caso contrário. Já no segundo caso , o elemento a_{ij} da Matriz A assume valor 1 caso exista uma conexão direcionada do vértice i ao vértice j e 0 caso contrário.

Cada tipo de representação, quando convertido em estrutura de armazenamento tem suas vantagens e desvantagens. A lista de conexões permite maior economia de memória (para redes esparsas) quando comparada com a matriz de adjacência, porém o acesso às arestas se torna mais complexo, pois se torna necessário buscar na lista. Dada a característica esparsa da rede analisada neste trabalho, a lista de conexões foi escolhida como estrutura de armazenamento.

2.4. Medidas de caracterização

A maior parte das redes reais é formada por milhares, ou mesmo milhões, de vértices e arestas. Dessa forma, uma inspeção visual de tais redes não se mostra suficiente para obter informações relevantes sobre a sua organização. Assim sendo, se torna necessário usar descritores de sua topologia, que são medidas de redes complexas.

As medidas topológicas vêm sendo desenvolvidas para análise, caracterização e classificação de redes complexas [2, 3, 6, 4]. Uma medida simples é a conectividade *ki* de um dado vértice *ni*, também conhecido como grau do vértice, que é igual ao seu número de conexões. Baseado na representação da matriz de adjacência, esta medida pode ser calculada conforme a equação 1. Além desta medida, temos também a conectividade média

<*k*>, que é usada para quantificar a densidade de conexões da rede, e pode ser calculada conforme a equação 2.

$$k_i = \sum_{j \in \mathcal{N}} a_{ij}. \tag{1}$$

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i \tag{2}$$

Para redes direcionadas, ou dirigidas, como é o caso da rede presente nesse trabalho, adota-se a medida de conectividade de um vértice dividida entre as conexões de entrada (*Kin*) e as conexões de saída (*Kout*). Baseado nesta medida, pode-se calcular *P(Kin)*, a probabilidade de selecionar aleatoriamente um vértice com grau *Kin* dentre todos os vértices do grafo e análogamente *P(Kout)*, a probabilidade de selecionar aleatoriamente um vértice com grau *Kout* dentre todos os vértices do grafo.

Outra medida que pode ser adotada é o nível de associatividade *r* da rede [7], definido na equação 3, onde a_{ij} é a fração de arcos da rede que conectam um nó do tipo *i* a um do tipo *j* e b_i e c_i são as frações de cada tipo de categoria que está ligada ao nó *i*. Esta medida assume valor *r = 0* quando a rede tem um perfil dissociativo e valor *r = 1* quando a rede tem um perfil perfeitamente associativo. Esta medida indica, para redes em que os nós podem ser separados em diferentes grupos, o tipo de ligação existente entre esses nós. Para redes associativas (ou perfeitamente associativas), nós de um mesmo grupo se conectam com maior frequência do que nós de grupos diferentes.

$$r = \frac{\sum_i a_{ii} - \sum_i b_i c_i}{1 - \sum_i b_i c_i} \tag{3}$$

A equação 3 satisfaz as regras definidas nas equações 4, 5 e 6.

$$\sum_{ij} a_{ij} = 1 \tag{4}$$

$$\sum_{j} a_{ij} = b_i \tag{5}$$

$$\sum_i a_{ij} = c_j \tag{6}$$

Além das medidas apresentadas, existe uma série de outras medidas que podem ser aplicadas na caracterização de redes complexas, entretanto as mesmas não foram empregadas no desenvolvimento deste trabalho devido à limitação computacional considerando o tamanho da rede analisada.

2.5. Considerações Finais

Neste capítulo procurou-se fornecer a base teórica necessária para compreender o trabalho realizado neste projeto. Foi apresentada a importância das redes complexas, seus conceitos básicos, representação e medidas de caracterização que serão empregadas durante o desenvolvimento do trabalho.

No próximo capítulo é descrito o desenvolvimento do projeto, discutindo-se cada etapa envolvida e o emprego dos conceitos abordados.

CAPÍTULO 3: DESENVOLVIMENTO DO TRABALHO

3.1. Considerações Iniciais

Devido a popularidade dos sites de e-commerce e a grande diversidade de produtos oferecidos por eles, torna-se necessário adotar estratégias de marketing para aumentar o consumo, dentre elas a exibição de produtos que frequentemente são comprados em conjunto.

Esse trabalho objetiva analisar o relacionamento entre diferentes categorias de produtos que são comprados em conjunto e caracterizar a rede formada entre esses produtos.

Nas próximas sessões será descrito todo o desenvolvimento do projeto para alcançar o objetivo proposto, detalhando a aplicação desde a etapa de captura de dados até a caracterização da rede formada.

3.2. Projeto

O principal objetivo deste trabalho é representar a relação entre os produtos frequentemente comprados em conjunto em um site de e-commerce e caracterizar a rede formada como resultado dessa representação. Além disso, deseja-se tirar conclusões sobre a organização dessa rede com base na caracterização obtida.

Para alcançar os objetivos do trabalho, o projeto foi dividido em 4 etapas simples e bem definidas, com objetivos claros e colocadas em sequência de tal forma que o resultado de uma etapa é usado como entrada da próxima etapa e, ao final da última etapa, o objetivo do trabalho seja alcançado.

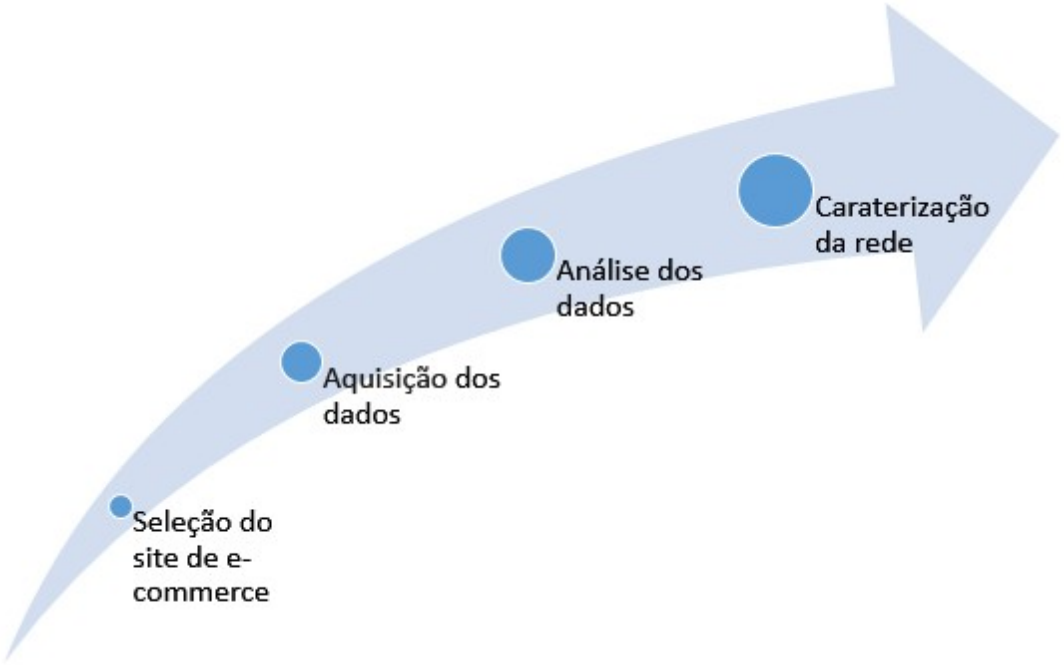


Figura 3: Etapas do projeto

Conforme a figura 3, pode-se ver a sequência de etapas que compõem o projeto. Cada uma das 4 etapas possui atividades específicas, as quais são detalhadas nas próximas sessões.

3.3. Descrição das Atividades Realizadas

As atividades realizadas neste projeto serão descritas a seguir, de acordo com a fase do projeto a qual se referem.

3.3.1. Seleção do site de e-commerce

Essa etapa do projeto é de suma importância, dado que é a base das demais atividades, sendo necessário escolher um site que forneça informações relevantes e que possibilite a extração das informações sem grandes barreiras.

Analisando o primeiro critério, na escolha de um site que forneça informações relevantes, o ideal é que seja selecionada uma fonte que permita agrupar os produtos em categorias, apresentando o valor e o nome do produto em questão. Além das informações

básicas, é necessário que o site apresente uma lista de produtos que frequentemente são comprados em conjunto para que baseado nessa informação seja possível criar uma rede.

Com relação ao segundo critério, na escolha de um site que possibilite a extração das informações sem grandes barreiras, espera-se que o site escolhido não imponha bloqueios à aquisição das informações, como por exemplo limitar o número de páginas visitadas ou a velocidade de download da página.

Baseado nos dois critérios apresentados, havia a opção entre o site brasileiro Livraria Saraiva[2] e o site americano Amazon[3], pois ambos fornecem informações relevantes ao estudo. Entretanto o site da Livraria Saraiva fazia uso de um sistema complexo para listar os itens frequentemente comprados em conjunto, enquanto o site Amazon fazia uso de um sistema mais simples com a mesma finalidade, assim, o site americano acabou sendo a escolha final.

A partir desta decisão, o site escolhido foi analisado e as informações pertinentes ao estudo foram identificadas, levando ao desenvolvimento de um conjunto de funções para permitir o avanço para a próxima fase.

2 http://www.livrariasaraiva.com.br

3 http://www.amazon.com

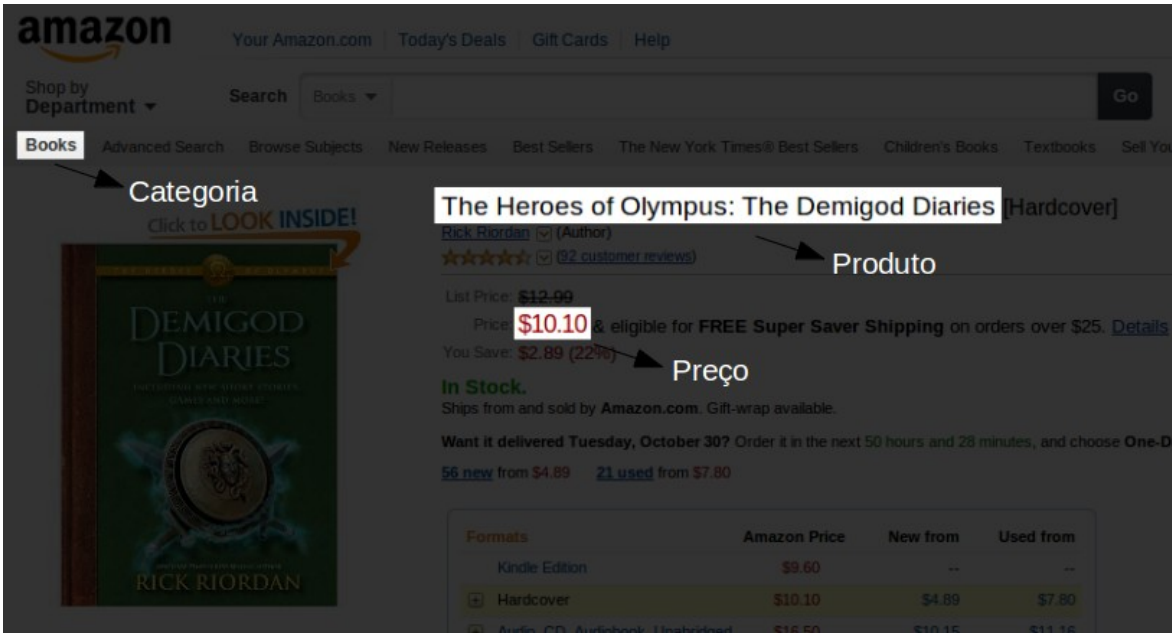


Figura 4: Informações extraídas da página.

Pode-se observar na figura 4, a disposição das informações relevantes que foram extraídas, em uma página qualquer do site escolhido. O resultado da extração desta página segue o formato V_1 = (“The Heroes of Olympus: The Demigod Diaries”, “Books”, 10.10).

Figura 5: Formação da rede entre os produtos que são frequentemente comprados juntos.

Na figura 5, é apresentada a relação entre o produto sendo visualizado e os demais produtos que costumam ser comprados em conjunto. Pode-se ver o esquema de formação da rede que liga os produtos, no qual a origem das setas indica o produto que está sendo minerado e o destino das setas indica os produtos relacionados a este. Essas informações são inseridas na base de dados como um grafo direcionado usando uma lista de conexões.

3.3.2. Aquisição de dados

Para esta etapa, foi adotada a linguagem de programação PHP dado a sua praticidade, velocidade de desenvolvimento e experiência prévia do aluno em relação à mesma. Além destes fatores, a presença de elementos como expressões regulares, funções nativas para download de páginas e conexão com um SGBD – Sistema Gerenciador de Bases de Dados –, tornam a linguagem uma boa escolha. Para armazenar os dados, foi utilizado o SGBD Oracle MySQL, por conta da facilidade de uso do mesmo e experiência prévia do aluno em relação a este.

Para a execução dessa fase, foi elaborado um script que realiza o download de uma página inicial fornecida pelo usuário, extrai as informações relevantes do produto (categoria, nome e preço), extrai as informações sobre relacionamento com os outros itens disponíveis na página e armazena todas essas informações em uma base de dados.

Cada novo item descoberto é armazenado em uma fila de espera para ser minerado posteriormente e o processo é repetido para todos os produtos presentes na fila, até que a rede formada tenha o tamanho desejado ou um dado intervalo de tempo desejado tenha terminado.

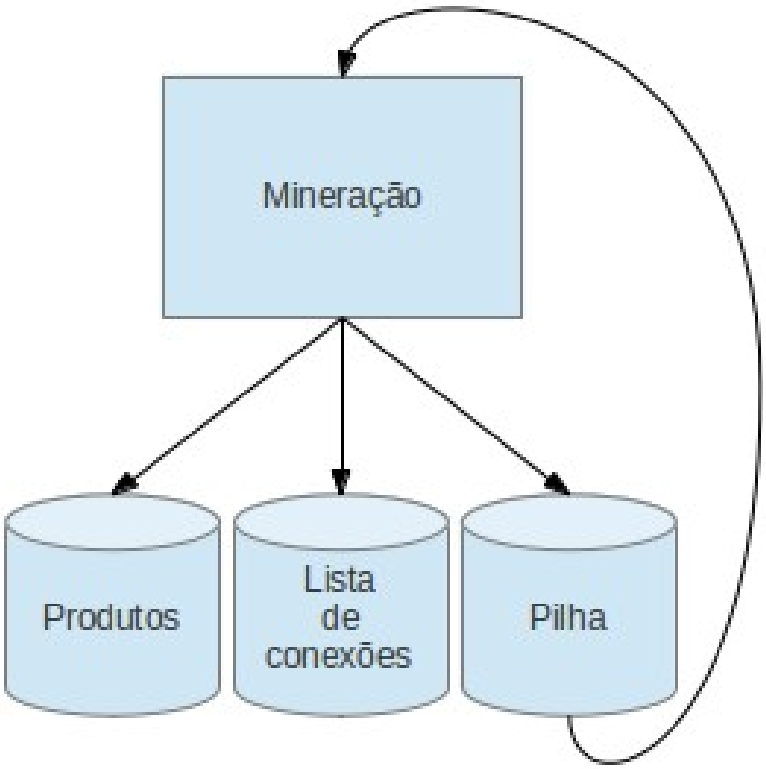


Figura 6: Esquema de funcionamento da extração de dados

A figura 6 destaca o funcionamento da principal atividade realizada durante esta etapa e os elementos presentes no ciclo de funcionamento do script elaborado para executar essa tarefa. A página sendo minerada tem os dados do produto (categoria, nome, preço), armazenados na base **Produtos**, as relações entre os produtos são armazenadas na base **Lista de conexões** e os novos produtos encontrados são armazenados na **Pilha**.

A tarefa de extração de dados foi iniciada a partir do livro “The Heroes of Olympus: The Demigod Diaries”, selecionado ao acaso. Ao término dessa etapa, foi obtida uma rede contendo 544.101 nós, 26.361.119 conexões entre esses nós e os produtos foram agrupados em 36 categorias, após a execução do script por cerca de 7 dias, com 15 tarefas em paralelo. É importante notar que cada produto pertence a apenas uma única categoria e

itens duplicados são evitados com base no nome do produto, desconsiderando-se detalhes como tipo de capa ou cor, por exemplo.

Categoria	Número de itens	Arcos de entrada	Arcos de saída
All Electronics	4062	159773	162402
Appliances	14	178	275
Arts, Crafts & Sewing	1670	65586	66276
Automotive	1037	28007	28827
Baby	846	34411	28024
Beauty	1756	49698	51216
Books	449887	22231002	22261130
Camera & Photo	512	13828	13803
Car Electronics	90	2031	2320
Cell Phones & Accessories	1193	56708	52577
Clothing & Accessories	75	1988	2088
Computers	1447	49685	53819
Furniture & Decor	6	52	59
GPS & Navigation	54	2444	1649
Grocery & Gourmet Food	2082	80759	75049
Health & Personal Care	2388	67405	64665
Home & Kitchen	3848	100035	108591
Home Improvement	2772	108006	102125
Industrial & Scientific	27	1101	1015
Jewelry	689	16559	16841
Kindle Store	340	11774	12636
Kitchen & Dining	4468	156759	159647
Magazine Subscriptions	354	20546	20650
MP3 Players & Accessories	429	12470	13846
Music	16474	929903	918949
Musical Instruments	1566	76343	73320
N/A	2479	62652	73755
Office & School Supplies	5411	227808	211838
Patio, Lawn & Garden	1151	34378	33779
Pet Supplies	551	19387	17442
Software	391	12832	12750
Sports & Outdoors	3532	115860	112286
Sports Collectibles	29	740	812
Toys & Games	28871	1438491	1445058
Video Games	3123	160330	150235
Watches	477	11590	11365

Tabela 1: Propriedades dos dados obtidos na etapa de aquisição de dados.

Na tabela 1 pode-se observar que a categoria “Books” possui uma quantidade de itens superior as demais categorias, isso ocorre pois a aquisição de dados foi iniciada a partir de um produto desta categoria e, conforme será apresentado, este é um

comportamento comum da rede. Diferentes produtos iniciais levam a diferentes redes com diferentes propriedades.

3.3.3. Análise dos dados

Para a etapa de análise dos dados, diversas consultas SQL foram criadas e usadas para sintetizar e extrair informações referentes às conexões entre os produtos, além de alguns scripts para possibilitar a transformação dos resultados das consultas em planilhas.

Além da síntese e extração das informações, durante esta etapa também foram calculados os valores da medida *Kin* e *Kout* para cada item e o nível de associatividade para a rede, obtendo-se o valor $r = 0.8301$ indicando que a rede possui um perfil associativo.

Categoria	Número de itens	Preço médio	Posição (entrada)	Posição (saída)
Magazine Subscriptions	354	$24.62	1	1
Music	16474	$14.81	2	2
Video Games	3123	$30.07	3	5
Toys & Games	28871	$19.71	4	3
Books	449887	$19.98	5	4
Musical Instruments	1566	$44.48	6	6
Cell Phones & Accessories	1193	$20.7	7	7
GPS & Navigation	54	$102.59	8	22
Office & School Supplies	5411	$14.66	9	10
Industrial & Scientific	27	$18.28	10	11
Baby	846	$23.53	11	17
All Electronics	4062	$72.53	12	8
Arts, Crafts & Sewing	1670	$16.21	13	9
Home Improvement	2772	$26.98	14	14
Grocery & Gourmet Food	2082	$19.77	15	15
Pet Supplies	551	$15.63	16	21
Kitchen & Dining	4468	$27.39	17	16
Kindle Store	340	$23.03	18	13
Computers	1447	$78.63	19	12
Software	391	$68.31	20	18
Sports & Outdoors	3532	$30.11	21	20
Patio, Lawn & Garden	1151	$25.8	22	24
MP3 Players & Accessories	429	$38.49	23	19
Beauty	1756	$15.32	24	25
Health & Personal Care	2388	$19.39	25	30
Automotive	1037	$12.82	26	29
Camera & Photo	512	$64.28	27	31
Clothing & Accessories	75	$17.81	28	28
Home & Kitchen	3848	$21.89	29	26
Sports Collectibles	29	$24.26	30	27
N/A	2479	$14.34	31	23
Watches	477	$82.01	32	34
Jewelry	689	$30.02	33	33
Car Electronics	90	$43.73	34	32
Appliances	14	$119.95	35	35
Furniture & Decor	6	$124.24	36	36

Tabela 2: Categorias ordenadas de acordo com a razão entre o número de arcos de entrada e o número de itens para cada categoria.

A tabela 2 apresenta a lista de categorias obtida, o número de itens presentes em cada categoria, o preço médio dos produtos e está classificada de acordo com a razão entre o número de arcos de entrada e o número de itens para cada categoria. A razão usada na classificação evita que o elevado número de itens de algumas categorias influencie diretamente na classificação, pois estas apresentariam um maior número de arcos de entrada.

Da tabela 2, observa-se que, em geral, as categorias que estão mais para o fim da lista possuem um alto preço médio dos produtos e/ou são categorias com produtos superfluos, como por exemplo jóias. O posicionamento das categorias na tabela destaca a popularidade da categoria dentro do site de comércio eletrônico.

3.3.4. Caracterização da rede

Nesta etapa, foram utilizadas todas as informações reunidas nas etapas anteriores a fim de caracterizar a rede e permitir que o comportamento de compra seja descoberto.

Gráfico 1: Histograma para o número de conexões de entrada.

O gráfico 1 relaciona o número de itens com o número de conexões de entrada, a partir disso, nota-se que existem poucos itens que são muito referenciados e portanto comprados em conjunto, enquanto a maioria dos itens é pouco referenciada. Esse comportamento se aproxima de uma lei de potência [8].

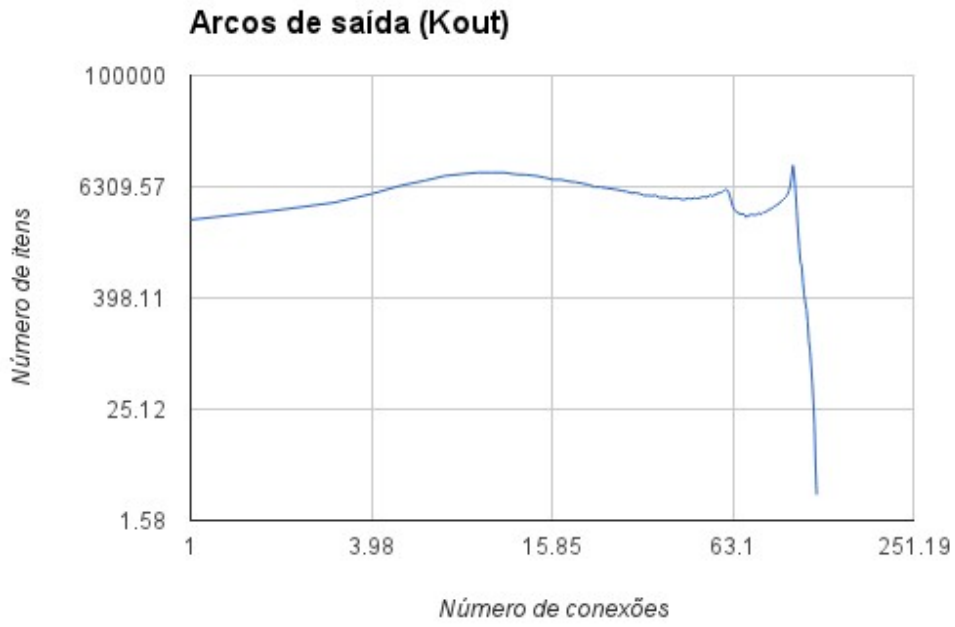
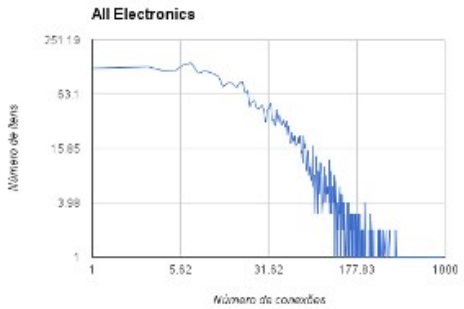


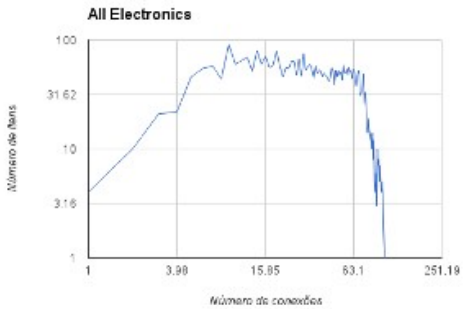
Gráfico 2: Histograma para o número de conexões de saída.

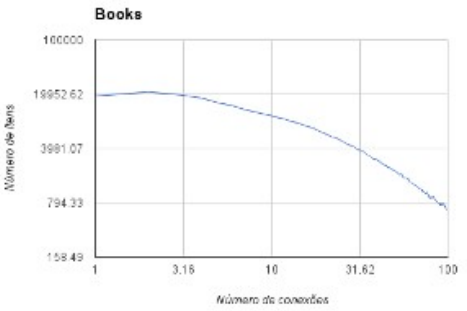
O gráfico 2, assim como o gráfico 1, relaciona o número de itens com o número de conexões, neste caso, de saída e a partir dele observa-se que as conexões de saída apresentam uma distribuição quase uniforme, não trazendo maiores conclusões para o estudo.

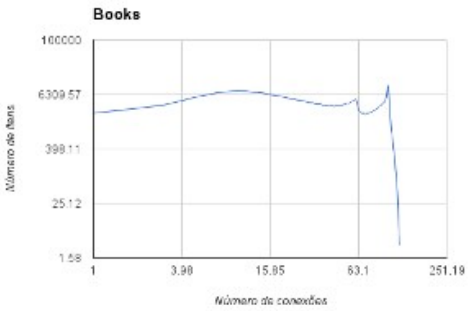
Conexões de entrada

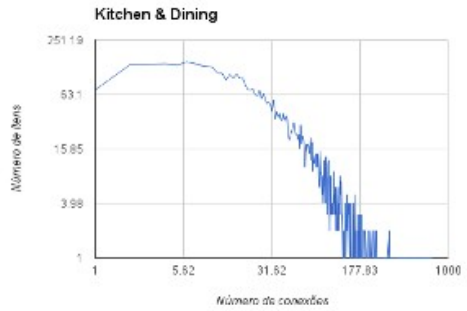


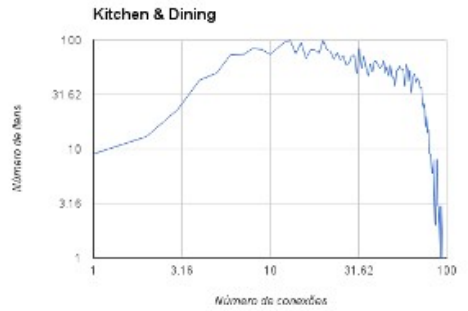
Conexões de saída

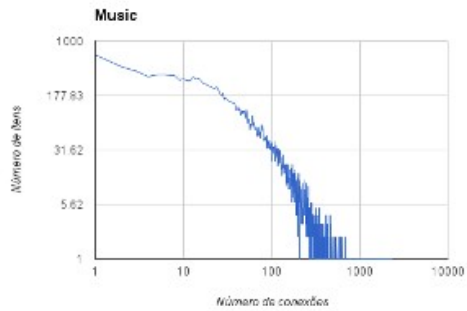


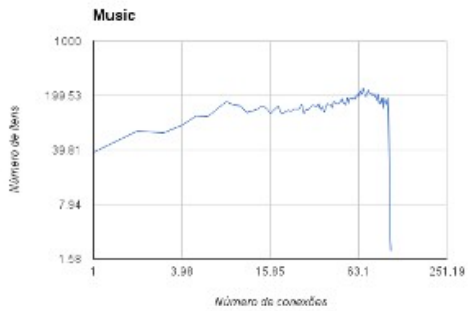

Books
100000
19952.62
3981.07
794.33
158.49
Número de itens
1
3.16
10
31.62
100
Número de conexões

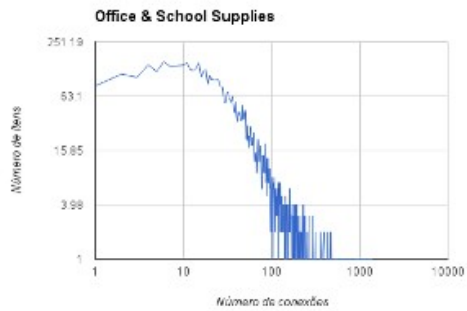

Books
100000
6309.57
398.11
25.12
1.58
Número de itens
1
3.98
15.85
63.1
251.19
Número de conexões

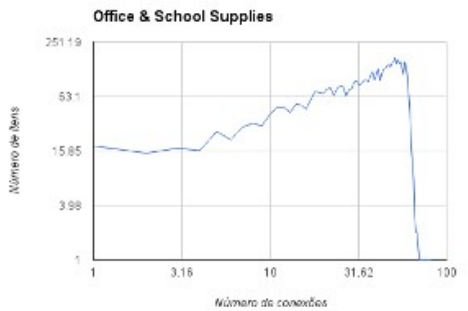

Kitchen & Dining
251.19
63.1
15.85
3.98
1
Número de itens
1
5.62
31.62
177.83
1000
Número de conexões


Kitchen & Dining
100
31.62
10
3.16
1
Número de itens
1
3.16
10
31.62
100
Número de conexões


Music
1000
177.83
31.62
5.62
1
Número de itens
1
10
100
1000
10000
Número de conexões


Music
1000
199.53
39.81
7.94
1.58
Número de itens
1
3.98
15.85
63.1
251.19
Número de conexões


Office & School Supplies
251.19
63.1
15.85
3.98
1
Número de itens
1
10
100
1000
10000
Número de conexões


Office & School Supplies
251.19
63.1
15.85
3.98
1
Número de itens
1
3.16
10
31.62
100
Número de conexões

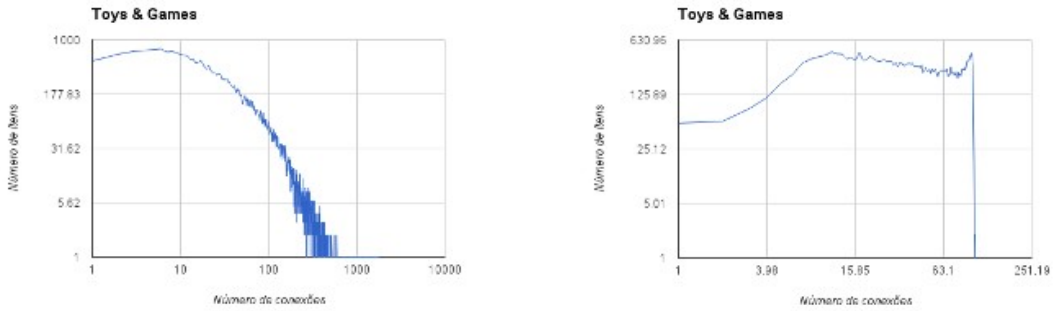


Tabela 3: Histograma de conexões de entrada e saída das 6 categorias com maior número de itens.

A tabela 3 apresenta os histogramas para as conexões de entrada e saída das 6 categorias com o maior número de itens na rede e a partir dela podemos observar que as categorias possuem distribuições de *Kin* e *Kout* semelhantes entre si e semelhantes a distribuição da rede. Nota-se também que a categoria “Books” possui uma distribuição de Kin levemente diferente das demais, provavelmente pelo número de itens que a categoria possui – é a categoria com o maior número de itens.

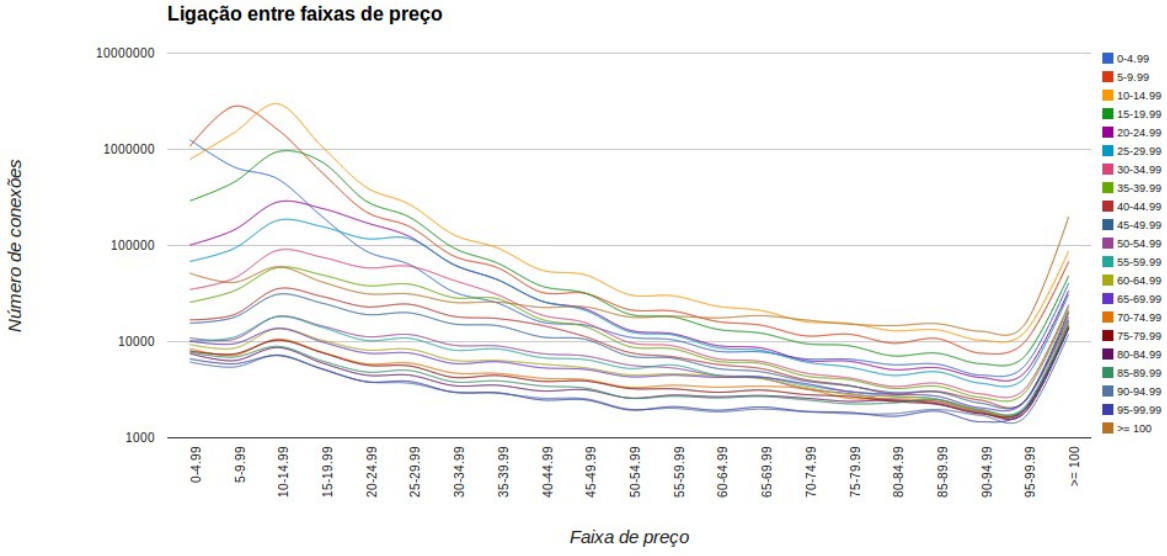


Gráfico 3: Ligação entre as faixas de preço.

O gráfico 3 apresenta a ligação entre faixas de preço, a partir do qual pode-se notar que os produtos na faixa de preços entre 0 e 25 dólares se relacionam fortemente com produtos nessa mesma faixa de preços. Além disso, também é possível notar que produtos com preço superior a 95 dólares se relacionam de maneira mais forte com produtos de valor superior a 100 dólares, mas também com produtos na faixa entre 0 e 25 dólares.

A partir da análise do gráfico 3 notam-se os seguintes comportamentos: se o cliente está visitando um produto com valor até 25 dólares, há maiores chances dele comprar mais um produto com valor até 25 dólares do que um produto com valor superior; se o cliente está visitando um produto com valor superior a 95 dólares, há chances de que ele compre mais um produto com valor superior a 95 dólares ou um produto com valor entre 0 e 25 dólares.

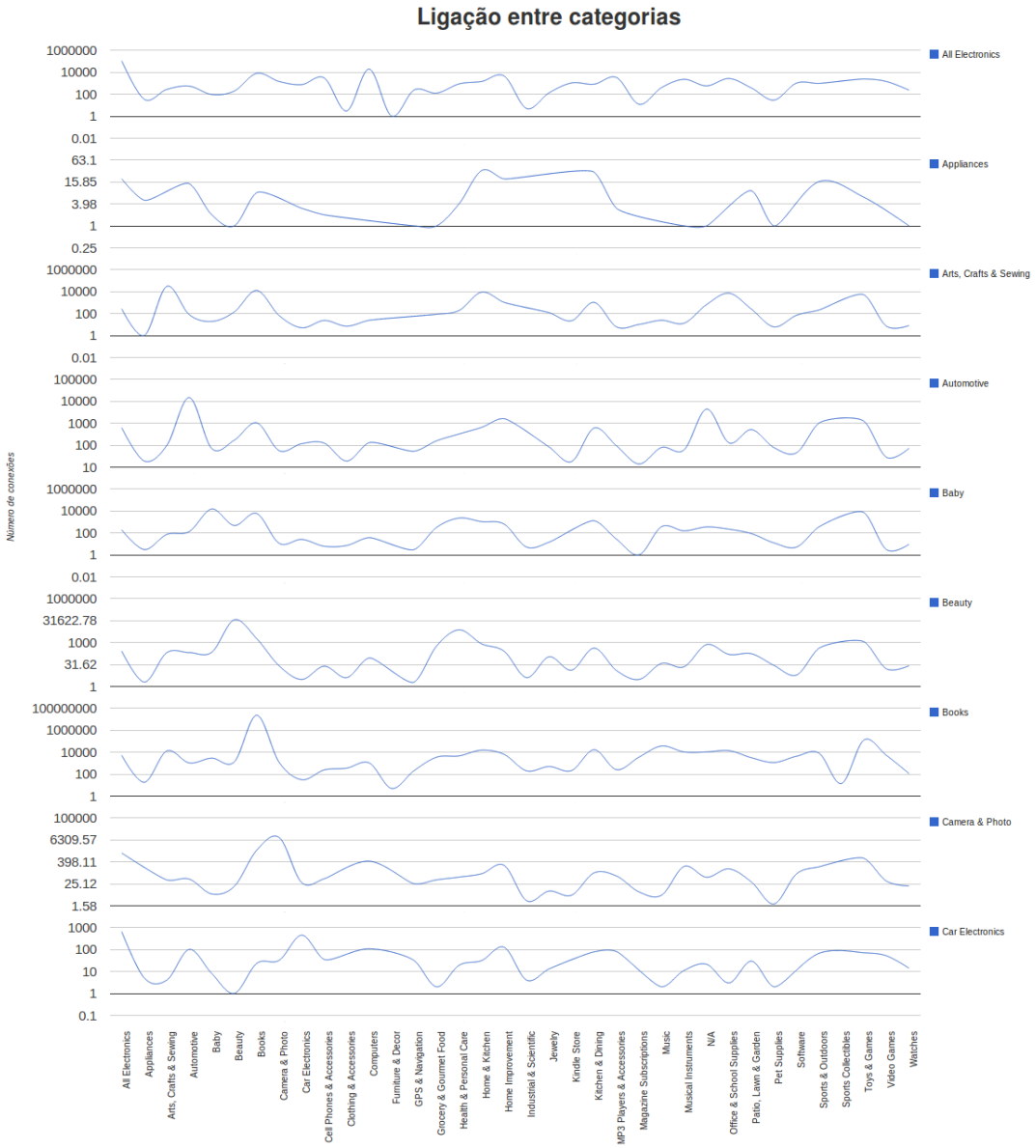


Gráfico 4: Ligação entre as categorias de produtos de "All Electronics" a "Car Electronics".

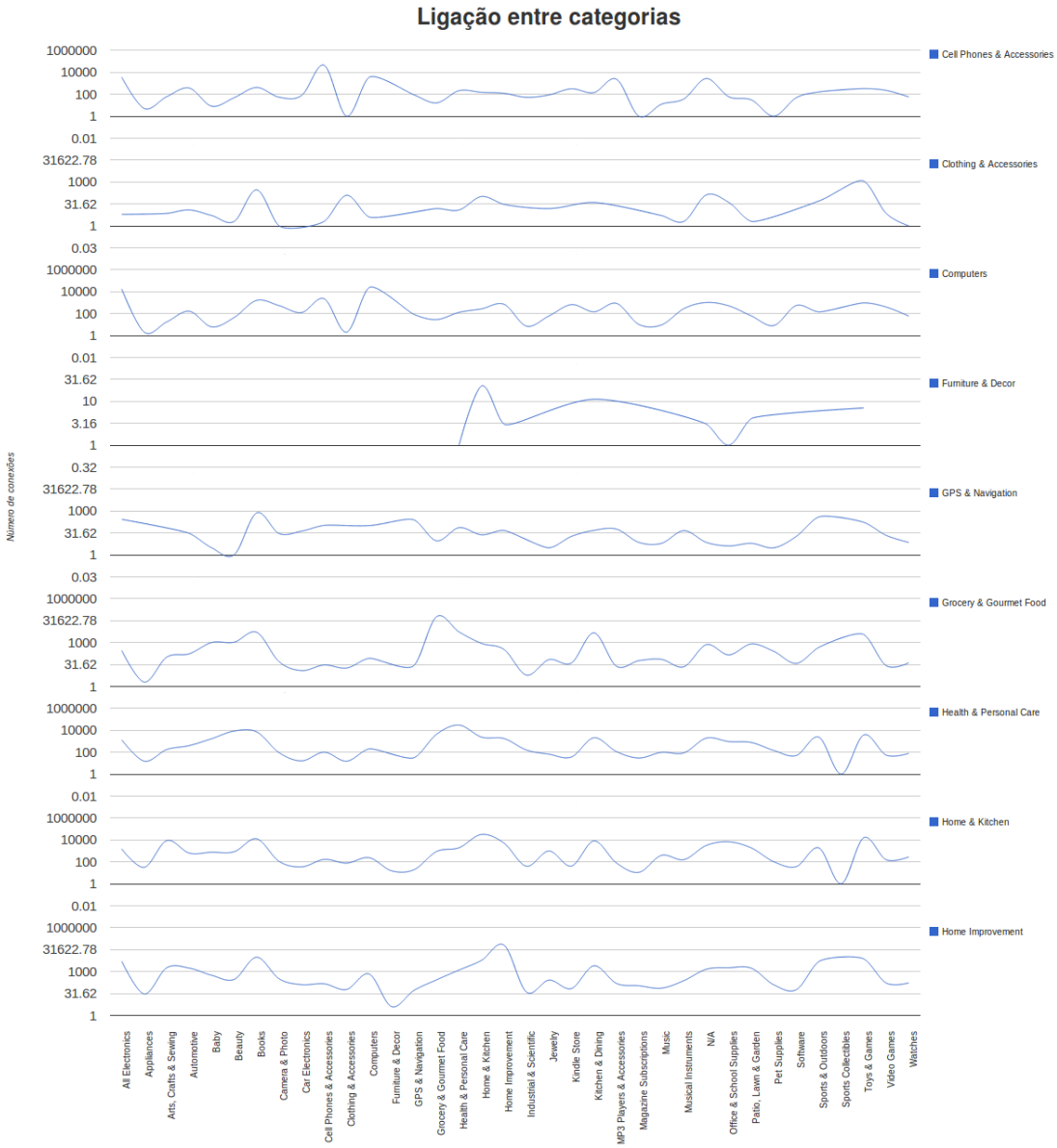


Gráfico 5: Ligação entre as categorias de produtos de "Cell Phones & Accessories" a "Home Improvement".

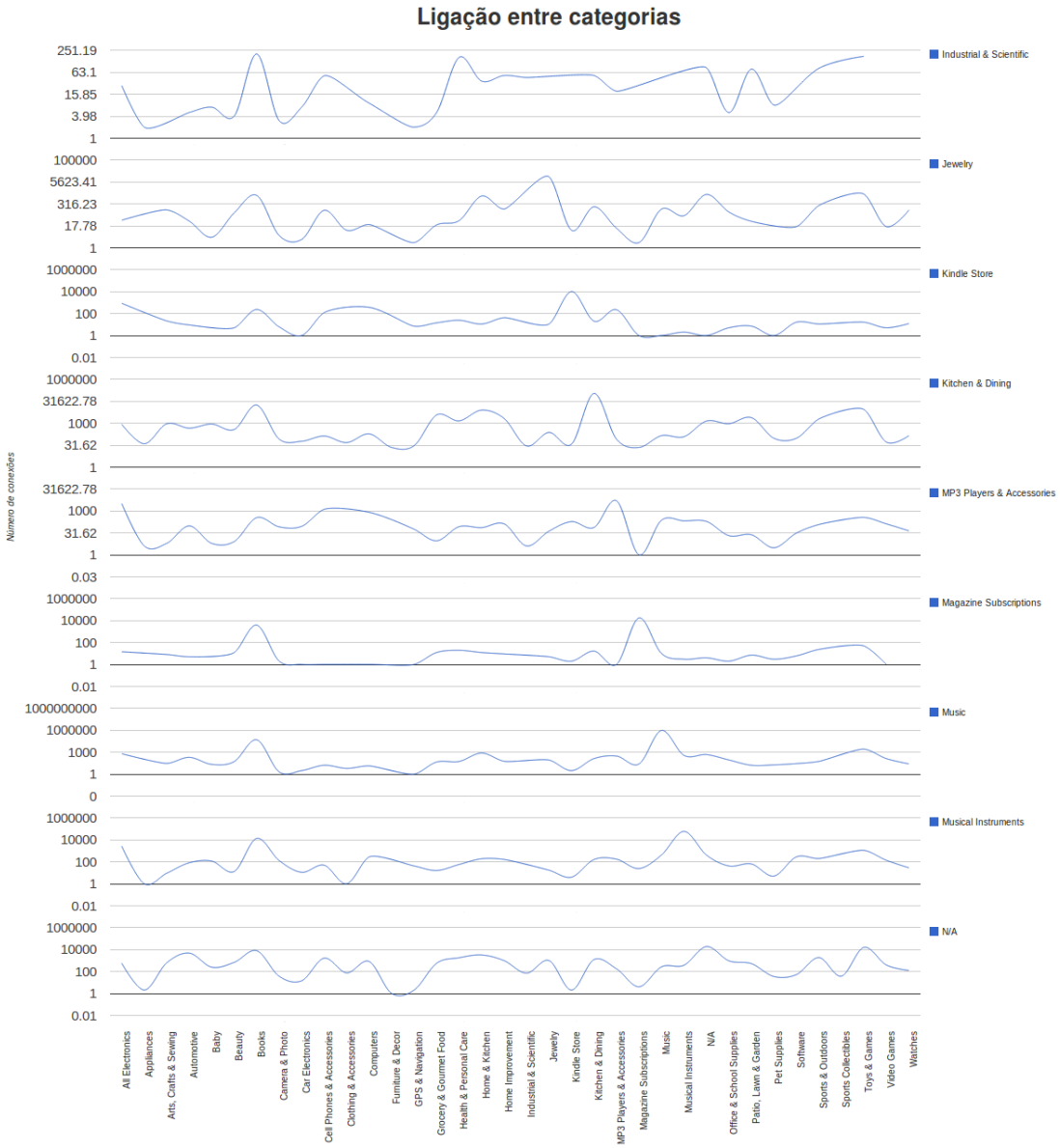


Gráfico 6: Ligação entre as categorias de produtos de "Industrial & Scientific" a "N/A".

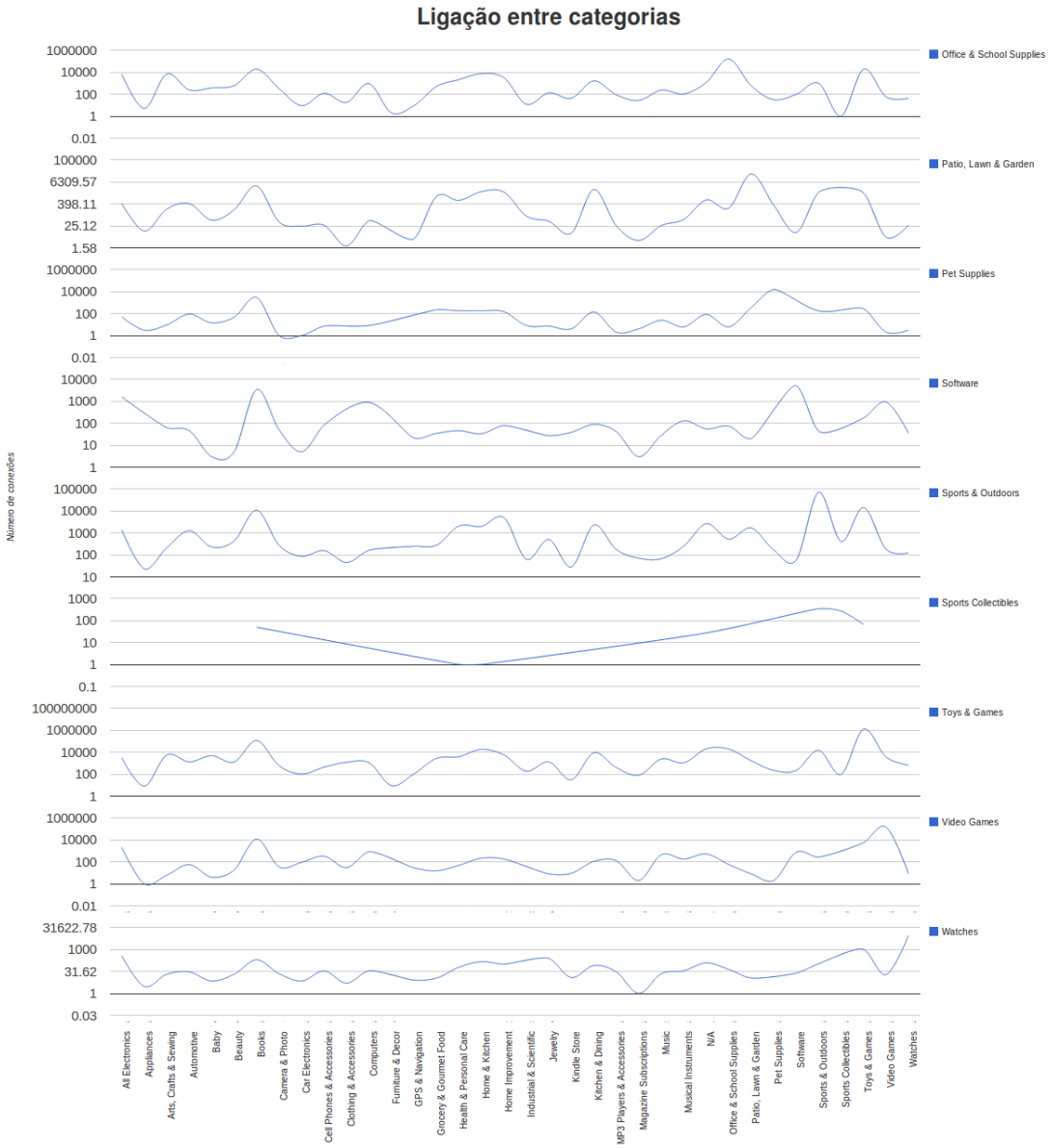


Gráfico 7: Ligação entre as categorias de produtos de "Office & School Supplies" a "Watches".

Os gráficos 4, 5, 6 e 7 apresentam o número de conexões entre as categorias e uma dada categoria, a partir dos quais pode-se observar, dada uma categoria, com quais outras categorias ela tem uma maior ou menor afinidade de conexão. Calculando o nível de associatividade da rede, foi obtido um valor de 0.8301, o que indica que a rede tem um perfil associativo, perfil este que pode ser observado nos gráficos 4, 5, 6 e 7, visto que em grande parte das categorias, a ligação entre os produtos da mesma categoria é muito maior

do que a ligação entre os produtos de categorias diferentes. Outro comportamento que pode-se observar é a ligação entre categorias similares, por exemplo itens da categoria “Car Electronics” tem um relacionamento forte com itens das categorias “All Electronics” e “Automotive”. Outro exemplo é a relação entre itens da categoria “Grocery & Gourmet Food” e o relacionamento com as categorias “Home & Kitchen” e “Kitchen & Dining”.

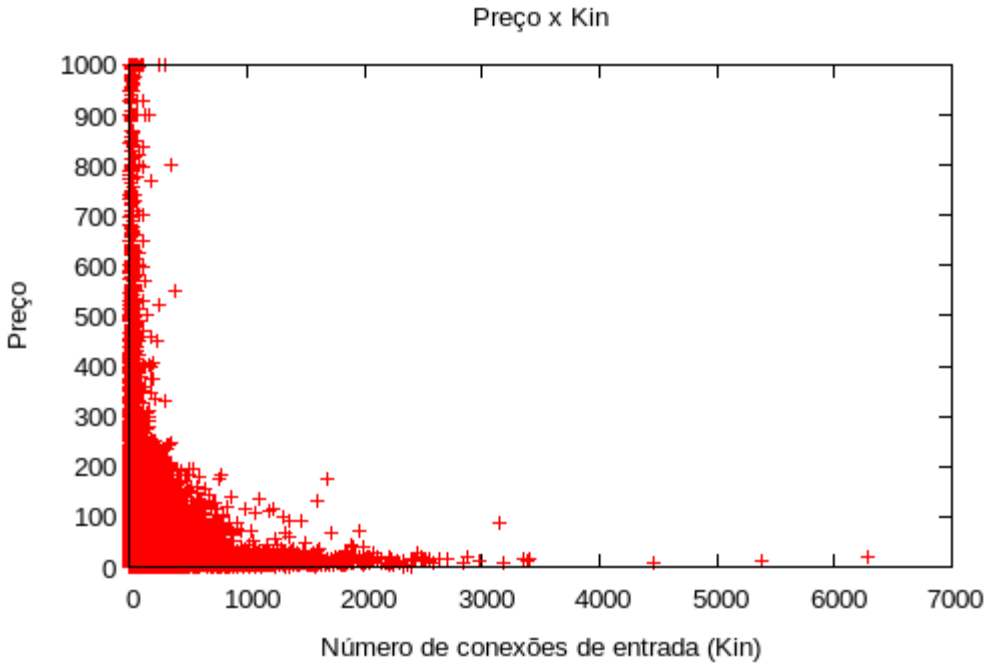


Gráfico 8: Relação entre o preço e o número de conexões de entrada.

O gráfico 8 apresenta a relação entre o preço dos produtos e o número de conexões de entrada, a partir do qual pode-se observar que produtos com elevados preços, possuem um reduzido número de arcos de entrada, logo, produtos caros tem uma tendência fraca a serem comprados em conjunto com vários itens, independente do valor desses itens. Entretanto, observa-se que produtos de menor preço, possuem um elevado número de arcos de entrada, assim, produtos baratos tem uma tendência forte a serem comprados em conjunto com vários outros itens.

3.4. Resultados Obtidos

Um dos primeiros resultados observados diz respeito a distribuição das conexões de entrada e saída dentro de cada categoria, nota-se que todas as categorias apresentam praticamente a mesma distribuição dentro da rede.

Analisando a rede conforme a faixa de preço dos produtos, observa-se que até 25 dólares, a tendência é que se compre outros produtos nessa mesma faixa de preço. Outro comportamento observado é o de que produtos com valores acima de 95 dólares levam a compra de produtos até 15 dólares e também produtos com valores acima de 95 dólares.

Sob a ótica das categorias, a rede tem um perfil associativo, ou seja, a tendência é que os produtos de uma mesma categoria se relacionem entre si de maneira mais frequente do que se relacionam com produtos de outras categorias. Além disso, também observa-se uma maior relação entre categorias semelhantes.

Um outro resultado observado é a popularidade de produtos de baixo preço, estes estão relacionados a vários produtos, já os produtos de alto preço não estão relacionados a muitos itens. Observa-se também que itens comuns, como assinatura de revistas, música, vídeo games e brinquedos são os mais populares. Assim, a popularidade dos produtos relaciona o valor e a utilidade do produto no dia a dia.

Por fim, o comportamento observado na rede é o mesmo comportamento encontrado na Internet e em Redes Biológicas [1], por exemplo.

3.5. Dificuldades, Limitações e Trabalhos Futuros

A maior dificuldade encontrada durante a execução desse trabalho foi relacionada a aquisição dos dados, visto que o site da Amazon usa AJAX para exibir a lista de produtos frequentemente comprados em conjunto e a requisição AJAX é baseada em diversas informações presentes na página, levando a um esforço para analisar e entender o funcionamento da página, de modo que fosse possível minerar os dados de maneira satisfatória. Além da dificuldade com o AJAX, por se tratar de um site de comércio eletrônico, a disponibilidade de estoque é dinâmica o que pode levar um item a estar indisponível no momento da aquisição, fazendo com que este item deixe de integrar a rede.

Além dos pontos levantados, o site da Amazon sofre alterações em sua interface e código ao longo da semana, invalidando parte das expressões regulares usadas na extração, forçando um monitoramento constante do processo de aquisição para garantir que o mesmo ainda estivesse funcional após alguns dias de execução.

Uma limitação muito grande encontrada foi o esforço computacional envolvido no trabalho com uma rede tão grande, levando ao aprendizado de estratégias alternativas para facilitar a obtenção de alguns resultados como por exemplo o uso de tabelas em memória para reduzir o tempo de varredura dos dados, além da contribuição para o conhecimento referente ao uso de base de dados e indexação de dados.

Por fim, o projeto teve sucesso em sua abordagem, visto que os resultados obtidos possibilitaram a caracterização da rede e levaram a um maior entendimento sobre o comportamento de compra. Para trabalhos futuros pode-se adotar um maior número de medidas para se extrair ainda mais informações sobre a rede, como por exemplo o nível de *clustering* da rede que não foi calculado, pois dado o tamanho da rede levaria muito tempo para ser calculado dados os recursos computacionais usados no desenvolvimento do trabalho.

3.6. Considerações Finais

Nesta seção especificou-se o desenvolvimento detalhado do projeto, com todos os seus sucessos e falhas. Foram detalhadas as fases do projeto, desde a escolha do site de e-commerce até a caracterização da rede obtida. Também foram detalhados os resultados obtidos em cada etapa.

Na próxima seção serão descritas as conclusões do projeto e da graduação na vida pessoal, profissional e acadêmica do aluno.

CAPÍTULO 4: CONCLUSÃO

4.1. Contribuições

Neste trabalho foram criados scripts que permitem a aquisição de dados oriundos de um site de comércio eletrônico e tais resultados podem ser usados no desenvolvimento de novos trabalhos.

Além disso, o desenvolvimento deste trabalho possibilitou o contato com problemas de processamento de Redes Complexas, permitindo o contato com a teoria de Redes Complexas que expandem e colocam em prática a formação acadêmica do aluno e o aprendizado de que existem diferentes abordagens que levam a um mesmo resultado.

Por fim, os resultados deste trabalho podem ser usados como critérios adicionais durante a criação da lista de produtos que serão oferecidos aos usuários durante uma visita. O emprego deste critério para a listagem pode levar ao aumento da probabilidade de uma compra conjunta.

4.2. Considerações sobre o Curso de Graduação

A graduação teve um papel importantíssimo na formação do meu caráter e da minha opinião e visão sobre o mundo. Além de me permitir ter um ponto de vista analítico e muitas vezes crítico, me mostrou novas formas de encarar a realidade.

Com a conclusão da graduação, percebo que parte das minhas maiores desilusões sobre a grade curricular eram em verdade prematuras, visto que algumas disciplinas ajudam a formar esse ponto de vista analítico e crítico que possuo atualmente. Entretanto continuo a advogar pela adoção de matérias que abordem assuntos mais relacionados à realidade do mercado. Acredito sim que a base oferecida pela universidade é sólida e tem sua importância, mas seria formidável ter à disposição matérias que tratassem dos assuntos, técnicas, padrões e ferramentas que estão em alta no mercado.

É incrível perceber como o tempo passou rápido, parece que foi ontem mesmo que eu estava em uma sala com mais um grupo de carecas, ouvindo a Profa. Renata falar sobre o curso. Hoje me sinto feliz em concluir uma etapa tão importante em minha vida.

REFERÊNCIAS

[1] L. da F. Costa, O. Oliveira, G. Travieso, F. A. Rodrigues, P. Villas Boas, L. Antiqueira, M. P. Viana, and L. Correa Rocha. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. Advances in Physics, 60(3):329-412, 2011.

[2] R. Albert and A. -L. Barabási. Statistical mechanics of complex networks. Reviews of Modern Physics, 74:48-98, 2002.

[3] M. E. J. Newman. Structure and function of complex networks. SIAM Review, 45(2):167-256, 2003.

[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chaves, and D. -U. Hwang. Complex networks: structure and dynamics. Physics Reports, 424:175-308, 2006.

[5] M. E. J. Newman and J. Park. Why social networks are different from other types of networks. Physical Review E, 68(3), 2003.

[6] L. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas. Characterization of complex networks: A survey of measurements. Advances in Physics, 56(1):167-242, 2007.

[7] M. E. J. Newman. Mixing patterns in networks. Physical Review E, 67, (2003).

[8] M. E. J. Newman. Power laws, Pareto distributions and Zipf's aw. Contemporary Physics, 46(5), 2005.