# Análise de variações estruturais em redes de personagens na série literária Harry Potter

**Eduardo de Mello Castanho[1], Patrícia de Andrade Kovaleski[1]**

[1]Escola Politécnica – Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brazil

{eduardocastanho,kovaleski}@poli.ufrj.br

***Abstract.*** *This study consist of analysing the behavior of a fictitious characters networks in a book series, evaluating its structural variations throughout the books and the impact caused by directed attacks. The network consists of characters as vertices and undirected weighted edges that represent the occurrence of adjacent characters in the book.*

***Resumo.*** *Este estudo consiste em analisar o comportamento de uma rede de personagens fictícios ao longo de uma série literária, avaliando suas variações estruturais ao longo dos livros e o impacto causado por ataques direcionados. A rede é composta por personagens como vértices e suas arestas não direcionadas com peso representam a ocorrência de adjacências de personagens no texto.*

## 1. Introdução

Este estudo foi inspirado em redes famosas estudadas ao longo da disciplina, combinando as ideias de adjacência entre palavras [Newman, 2006] e ocorrência de personagens em um capítulo [Knuth, 1993]. Acreditamos que avaliar a adjacência de personagens no texto produz uma rede mais fiel à ideia de relacionamento entre personagens e resulta em um grafo menos denso e mais elucidativo.

A principal motivação é a de avaliar o comportamento da rede ao longo dos livros, tentando buscar semelhanças com o que se espera em redes reais de pessoas no mundo atual. As redes fictícias possuem as mesmas propriedades das redes sociais? Ou teriam um comportamento muito distinto? Por se tratar de um livro fantástico, com personagens irreais, mágicos, as relações podem diferir da realidade atual.

## 2. Artigos Relacionados

Redes de obras literárias podem ser construídas de diversas maneiras, desde avaliação de palavras e maneira como elas se relacionam, até análises elaboradas de diálogos e avaliações de sentimento dos personagens.

Um estudo técnico foi feito por Newman a partir de uma rede de adjacências de palavras na obra de *Charles Dickens*: "David Copperfield" [Newman, 2006]: cada palavra no texto é associada às palavras adjacentes, criando arestas. Neste estudo, Newman procura identificar estruturas de comunidade e centralidade de vértices em redes reais, dando ênfase aos seus algoritmos de forma mais abrangente, não se atendo à

natureza literária da rede. Sua principal preocupação é aplicar o algoritmo em uma rede real muito densa, como é a de adjacências de palavras em um texto.

Um problema mais relacionado à literatura foi discutido na construção da rede baseada na obra de Victor Hugo: "Les Miserables" [Knuth, 1993]: a identificação de personagens em obras literárias. Em sua rede, cada vértice é um personagem e cada aresta representa uma menção de dois personagens em uma mesma página. Em seu livro, Knuth relata diversas dificuldades encontradas na caracterização da rede, sendo necessário um trabalho manual não automatizável para garantir um mínimo de confiabilidade à rede. Ele ressalva, porém, que sua rede possui imprecisões que somente com uma inteligência artificial muito eficiente seria possível resolver de forma automatizada.

Um artigo mais recente de Waumans et al. elaborou um estudo muito aprofundado sobre diálogos e análise de sentimentos dos personagens em diversas obras literárias [Waumans, 2015]. A rede é formada por personagens e as arestas representam diálogos entre dois personagens no livro. O maior desafio foi o de identificar interlocutores, uma vez que não é trivial para um algoritmo entender a quem se dirige um diálogo no livro. Em sua análise, os autores focaram na série literária de J. K. Rowling, "Harry Potter", estudando sua estrutura e comparando com outras 40 obras literárias de relevância internacional.

## 3. Geração das Redes

### 3.1. Algoritmo para geração dos grafos

A ideia do algoritmo consiste em, com base nos textos literários, gerar uma rede de todos os personagens de uma obra literária como vértices e suas arestas representam ocorrências adjacentes de dois personagens.

Utilizamos dois tipos de entrada: arquivos txt de cada edição do Harry Potter original em inglês, disponível na *web*, e uma página da Wikipedia contendo 192 personagens da série.

O algoritmo inicialmente identifica todos os personagens presentes na página da Wikipedia e gera um arquivo txt em que cada linha representa um personagem, em ordem alfabética, associado ao seu índice, que é utilizado como identificador do vértice no grafo.

Em seguida, é lida cada linha de cada livro txt, identificando as palavras que começam com letra maiúscula e verificando se ela se encontram na lista de personagens. Caso haja apenas uma ocorrência dessa palavra na lista de personagens, compara-se o seu índice com o índice do último personagem encontrado (na primeira iteração, essa etapa é pulada). Caso os índices sejam diferentes, verifica-se se já existe uma aresta entre eles. Em caso positivo, o grau da aresta é incrementado; em caso negativo, a aresta é criada com grau um. O índice do vértice é então armazenado em uma variável referente ao último personagem encontrado.

Caso haja múltiplas ocorrências da palavra na lista de personagens, significa que há um conflito (por exemplo, sobrenomes iguais). Nesse caso, todos os índices conflitantes são salvos em uma lista e passa-se para a próxima palavra. Caso esta

também seja maiúscula e se encontre na lista de personagens, procura-se na lista a ocorrência de ambas as palavras em uma mesma linha, de forma a encontrar um único personagem com ambos os nomes.

Em último caso, se o conflito não for resolvido, as palavras são avaliadas individualmente e prevalece o personagem com maior grau, com critério de desempate sendo o índice do vértice.

### 3.2. Dificuldades

Como já relatado em artigos similares [Knuth, 1993], [Waumans, 2015], a identificação de personagens e suas relações não é algo trivial para um máquina. Discernir quais palavras representam os personagens e saber diferenciar aqueles com nomes similares ou da mesma família pode ser ambíguo até para um leitor comum.

No caso da rede do livro Harry Potter, resolvemos partir de uma lista de personagens disponível na internet. Porém, verificar ocorrências destes nomes no livro pode se tornar ambíguo pois muitas vezes os personagens são chamados somente pelo sobrenome, sendo então confundidos com seus parentes. Não há uma forma determinística de se saber com 100% de certeza qual é o personagem em questão, portanto resolvemos adotar a regra de que o personagem de maior grau têm prioridade no conflito, seguindo uma variação da regra de *preferential attachment*. Decidimos pela escolha direta pelo personagem de maior grau ou invés da aplicação de uma probabilidade para obtermos sempre as mesmas redes, facilitando comparações futuras. É evidente que isso implica em imprecisões, porém de uma forma geral atende as expectativas.

Outro problema seria encontrado em caso de primeiro nome igual entre múltiplos personagens, porém, felizmente, isso quase não ocorre na série da J. K. Rowling. Houve um caso em particular, porém, que gerou um grande problema: existem dois personagens, pai e filho, com exatamente o mesmo nome (Bartemius Crouch), variando apenas no acréscimo de Jr. ao filho. Para piorar a situação, ao longo do livro o filho raramente é chamado de Jr., sendo possível diferenciá-lo somente através do contexto. Não encontramos uma solução para esse caso e optamos por remover o filho da lista, uma vez que ele aparece muito pouco no livro, enquanto que o pai tem um papel significante.

Por fim, existe a questão de nomes ou sobrenomes que existem no dicionário, sendo também utilizados como palavras comuns (ex: Sirius Black). Nesse caso, mesmo diferenciando maiúsculas de minúsculas, a palavra pode ser confundida com o personagem quando aparecer no início de uma frase. Não tivemos soluções para esses casos, portanto os personagens com nomes dessa natureza tiveram um certo favorecimento.

## 4. Características da Rede

A rede é composta por vértices que representam os personagens do livro e suas arestas são formadas de acordo com a ocorrência adjacente de dois personagens ao longo do texto. As arestas não são direcionadas, porém possuem um peso referente à quantidade de ocorrências desta mesma aresta ao longo da série.

### 4.1. Propriedades Gerais

A rede total gerada a partir de todos os 7 livros do Harry Potter possui 192 vértices e 2724 arestas, o que resulta em um grau médio de 28,4. Trata-se de uma rede conexa muito densa (0.15), tendo como seu principal vértice o protagonista da série, com grau 186 (97% dos vértices). Com isso, naturalmente as distâncias mínimas são baixíssimas, pois praticamente todos os vértices estão conectados ao principal. Verifica-se, então, que a distância mínima média da rede é de 1.88, sendo que nenhum vértice têm distância mínima média acima de 3. De fato, o diâmetro é igual a 4 e o raio igual a 2.

A clusterização local média da rede é muito elevada, 0.66, o que indica que personagens relacionados têm alta probabilidade de se relacionaram com os vizinhos de seus vizinhos. Isso também é observado em redes reais de pessoas, como o Facebook [Ugander, 2011].

### 4.2 Arestas com peso

Para avaliar a intensidade da relação entre dois personagens, observamos o comportamento do peso da aresta entre eles. Das 2724 arestas, a de maior grau é a que une Harry Potter e seu melhor amigo Rony Weasley, atingindo um grau de 4524. As principais relações se deram de fato entre os personagens principais: Harry, Rony, Hermione, Dumbledore e Voldemort.

É interessante notar que a aresta que liga os irmãos gêmeos Fred e George Weasley possui o 10º maior grau, igual a 562, sendo a segunda maior aresta que não inclui o protagonista (a primeira é dada entre Rony e Hermione).

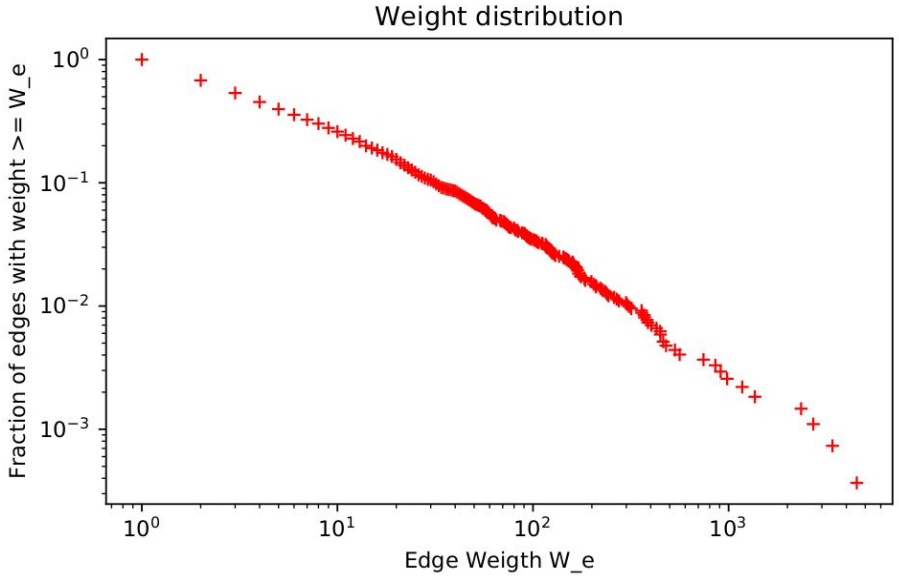


**Figura 1. Distribuição de grau das arestas**

O importante é observar que essa distribuição de grau segue uma lei de potência, como pode ser visto no gráfico 1: a maioria das arestas possuem grau menor do que 4, sendo a média é de 21.9, enquanto que a maior aresta, como dito anteriormente, possui grau 4524: mais de 200 vezes a média. Esse é mais um indício de que a rede em questão se assemelha a redes sociais reais.

### 4.3 Robustez

Para avaliar a robustez da rede, submetemos-na a uma série de ataques direcionados, de forma gulosa sobre o maior grau. Por haver vértices de grau muito alto, poderia-se imaginar que os ataques iriam surtir muito efeito, rapidamente desconstruindo a rede. O que aconteceu, porém, foi o contrário: por haver uma densidade muito alta, a rede conseguiu suportar bem os ataques.

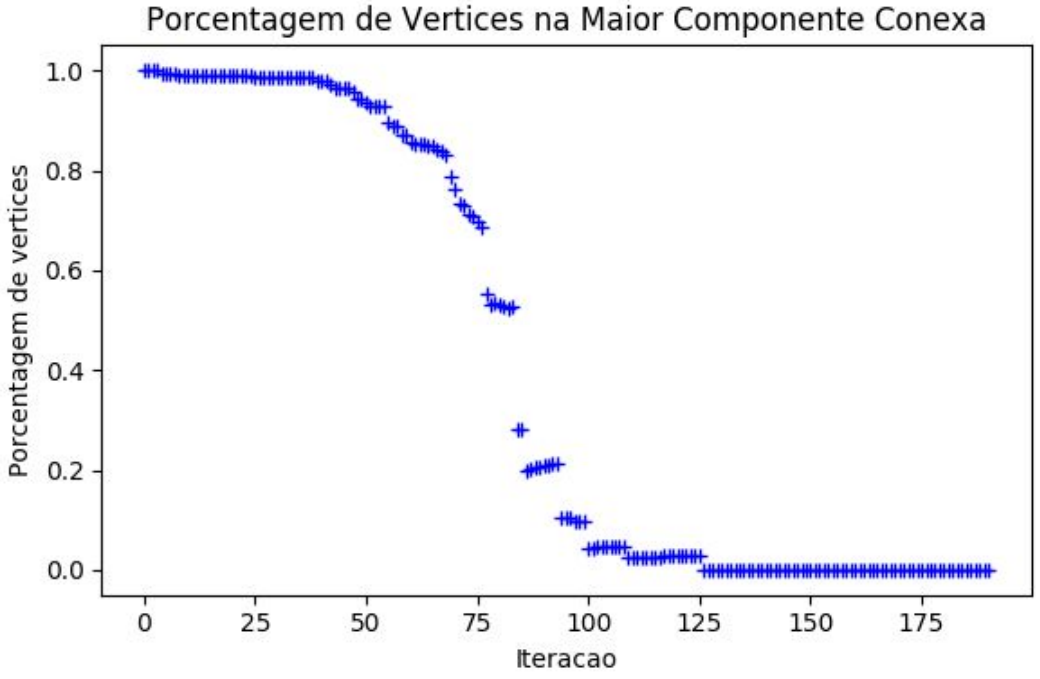


**Figura 2. Tamanho da maior componente conexa ao longo das iterações do ataque**

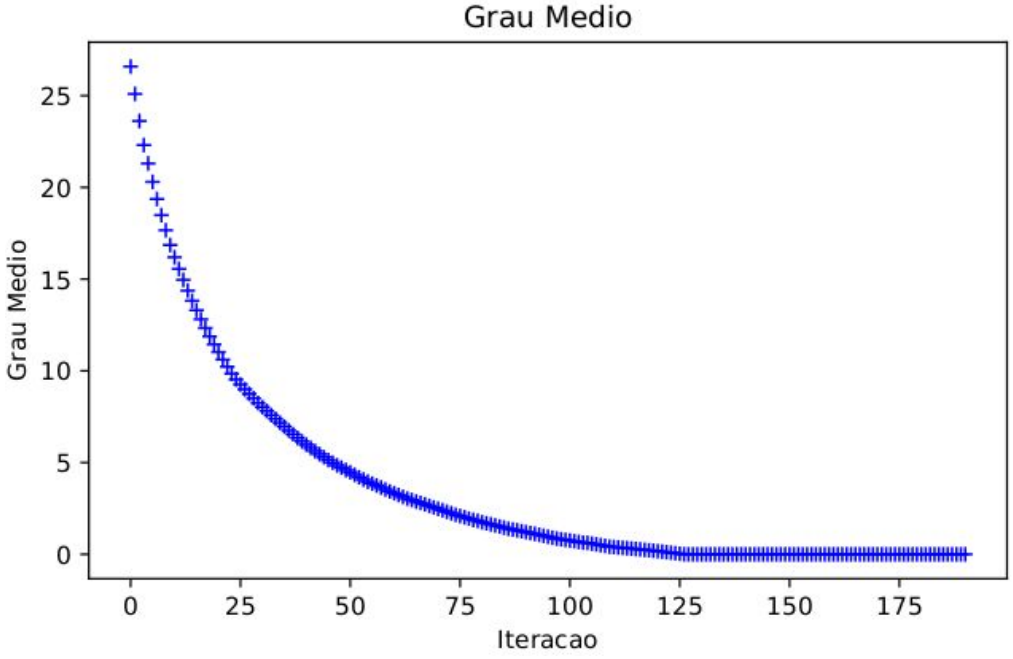


**Figura 3. Comportamento do grau médio ao longo das iterações do ataque**

Na Figura 2, podemos observar que a rede se mantém bem estruturada até aproximadamente a 75ª iteração (39% de vértices eliminados), tendo uma componente conexa ainda muito grande, próxima ao número total de vértices naquele instante. Só então que sua maior componente conexa começa a se desmantelar, até ficar praticamente inexistente na 100ª iteração. A rede então deixa de existir por completo por volta da 125ª iteração.

A Figura 3, por outro lado, indica um forte decréscimo do grau médio já nos primeiros ataques, seguindo um comportamento exponencial, como era de se esperar: eliminando os principais vértices, praticamente todos os outros vértices perdem grau pois estão muito próximos dos protagonistas. Uma rede com raio e diâmetro tão baixos inevitavelmente perde rapidamente o grau médio diante de ataques dessa natureza.

## 5. Evolução da Rede ao Longo da Série

Para avaliar a evolução da rede ao longo da série, geramos 7 redes distintas, que acumulam cada um dos livros a começar pelo primeiro: a primeira rede é gerada a partir do texto do primeiro livro; o segundo grafo agrega do segundo ao primeiro livro, e assim sucessivamente.

Na tabela 1, podemos acompanhar a evolução das propriedades do grafo conforme a rede vai crescendo ao longo da série. É interessante observar que, quanto mais nós vão sendo adicionados à rede, mais as propriedades aumentam: enquanto os nós aproximadamente são duplicados do primeiro para o sétimo livro, as arestas são quintuplicadas, o grau médio é triplicado e a densidade aumenta em 50%. Por outro lado, de forma intuitiva, a média das distâncias mínimas da rede diminui.

**Tabela 1. Evolução das propriedades da rede ao longo da série**

| Livros | Nós | Arestas | Grau Médio | Densidade | Média Distâncias Mínimas |
|---|---|---|---|---|---|
| 1 | 90 | 422 | 9,38 | 0,105 | 2,125 |
| 1-2 | 104 | 690 | 13,27 | 0,129 | 2,040 |
| 1-3 | 118 | 994 | 16,85 | 0,144 | 1,984 |
| 1-4 | 138 | 1448 | 20,99 | 0,153 | 1,927 |
| 1-5 | 162 | 1998 | 24,67 | 0,153 | 1,907 |
| 1-6 | 180 | 2321 | 25,79 | 0,144 | 1,911 |
| 1-7 | 192 | 2724 | 28,38 | 0,149 | 1,890 |

Esse fenômeno de aproximação dos vértices com o aumento da rede é contra-intuitivo em um primeiro momento, porém já foi evidenciado empiricamente nas mais diversas redes sociais do mundo real. Novamente, então, notamos a semelhança da rede de personagens fictícios com as redes reais.

## 6. Conclusão

Obras literárias, por mais fictícias que sejam, representam de forma geral relações humanas semelhantes à sociedade contemporânea à obra. Através desse estudo, pudemos identificar as semelhanças que as redes sociais fictícias possuem com as redes reais do mundo moderno. A robustez, a alta proximidade entre os vértices, os caminhos curtos para atingir qualquer vértice da rede, o aumento da densidade e estreitamento da rede em função do crescimento de nós da rede, lei de potência no grau das arestas, são todas características que encontramos nas redes como Facebook e Linkedin que também foram observadas na rede social do Harry Potter, obra contemporânea às redes digitais.

Talvez a maior disparidade encontrada na rede estudada é a alta concentração de vértices e extrema proximidade entre os personagens, porém isso é perfeitamente justificável por se tratar de uma história que gira em torno de alguns personagens principais, muito recorrentes ao longo de todo o livro. Um estudo realizado sobre as propriedades da rede do Facebook [Ugander, 2011] concluiu que, apesar de a rede como um todo ser muito esparsa, existem inúmeros subgrafos altamente conectados, com densidade elevada. Dessa forma, ao avaliar a rede do "Harry Potter", em se tratando de um mundo fechado, restrito a poucos personagens que frequentam os mesmo ambientes, é natural que as relações sejam muito mais estreitas do que no mundo real.

## Referências


Newman, M. E. J. "Finding Community Structure in Networks Using The Eigenvectors of Matrice" In: Phys. Rev. E pag. 74, 036104 (2006).

Knuth, D. E. "The Stanford GraphBase: A Platform for Combinatorial Computing" In: Addison-Wesley, Reading, MA (1993).

Waumans, Michaël C., Thibaut Nicodème, and Hugues Bersini. "Topology analysis of social networks extracted from literature." In: PloS one 10.6 (2015)

Ugander, J., Karrer, B., Backstrom, L., Marlow, C. "The Anatomy of the Facebook Social Graph" In: https://arxiv.org/pdf/1111.4503.pdf (2011)