Revista Brasileira de Ensino de Física, v. 27, n. 2, p. 389 - 393, (2005)
www.sbfisica.org.br

Uma aplicação da teoria de redes à estilometria: Comparando Machado de Assis e Tribuna do Norte

(An application of network theory to stylometry: comparing Machado de Assis and a local newspaper)

Gilberto Corso[1], Camilla R. Fossa e Genilson B. de Oliveira

Departamento de Biofísica e Farmacologia, Universidade Federal do Rio Grande do Norte, Natal, RN, Brasil
Recebido em 24/1/2005; Aceito em 4/2/2005

Este trabalho tem como objetivo ilustrar a teoria de redes através de uma aplicação no estudo quantitativo do estilo de textos, a estilometria. Construimos uma rede a partir das frases do texto e da posição relativa das palavras dentro das frases. Usamos como vértices da rede as frases e estabelecemos conexões entre os vértices cada vez que a mesma palavra aparece na mesma posição da frase. Comparamos neste trabalho textos de Machado de Assis, Rui Barbosa e fragmentos extraídos de um jornal local de nossa cidade: a Tribuna do Norte. Estimamos a dispersão na curva do número de conexões dos vértices da rede a qual se caracteriza por apresentar descontinuidades. Utilizamos esta grandeza estatística para caracterizar a riqueza estilística do texto.
Palavras-chave: teoria de redes, estilometria, lingüística.

The purpose of this paper is to illustrate the network theory through an application of quantitative studies of text styles. We build a network from sentences of a text and the relative positions of the words in the sentence. The sentences are used as vertices of the network and connections were established between the vertices each time the same word appeared in the same position in the sentence. We compare here, texts of Brazilian authors Machado de Assis and Rui Barbosa and articles of a local newspaper, Tribuna do Norte. We estimate the dispersion, D, of the curve of number of connections of the network vertices which is characterized by presenting strong discontinuities. The quantity D is used to characterize the stylistic richness of the text.
Keywords: networks, stylometry, linguistics.

1. Introdução

Este artigo é destinado a dois públicos distintos, sendo o primeiro formado por físicos e estudantes que desejam aprender um pouco mais sobre teoria de redes. Este grupo encontrará aqui um exemplo inusitado de rede e através dele poderá instruir-se e informalmente refletir sobre o assunto. Essencialmente a estas pessoas este trabalho é escrito, justificando assim um dos objetivos da revista: manter um espaço para divulgação de temas contemporâneos de Física. O segundo grupo dos eventuais interessados neste trabalho, mais seleto, é formado por pessoas envolvidas com estilometria: uma área da linguística que estuda técnicas matemáticas que avaliam o estilo de textos e de autores.

A Física nas últimas décadas tem passado por uma diversificação de interesses e objetos de estudo. Há um século esta ciência estudava a mecânica, a estrutura da matéria ou as transformações da energia. Hoje, físicos se aventuram nas mais inusitadas áreas conquanto possam se utilizar de métodos que lhe sejam familiares. Neste novo momento, as ciências da vida, a economia, ou até mesmo a linguística passaram a se tornar atraentes aos olhos de um tipo de cientista até então habituado a soluções exatas. Uma das surpresas desta *nouvelle vague* da física contemporânea tem sido a teoria de redes, pois ela surge fruto da especulação de físicos trabalhando fora dos temas clássicos de sua ciência. As redes passam a atrair a atenção desta comunidade na sociologia, computação ou biologia celular. Apesar de matemáticos estarem estudando redes há muito tempo, só na virada do século XX uma parcela da comunidade da física estatística voltou seus olhos para este campo. Em parte o recente interesse pelo assunto coincidiu com o advento da grande rede que conecta computadores do mundo inteiro - a Internet. Por outro lado, as redes têm recebido admirável atenção por se tratar de uma abordagem que se pressente seja fundamental em uma teoria, inexistente ainda, da complexidade. Mas independentemente do que seria a dita

[1]E-mail: corso@dfte.ufrn.br.

complexidade (não é nem vagamente do interesse dos autores explicitarem este conceito), ou dos motivos que tenham levado físicos a dirigirem seus interesses ao mundo das redes, o fato é que elas chegaram para ficar e hoje a teoria de redes se tornou um tópico da física estatística.

Uma rede é um objeto matemático bastante rudimentar quando comparado a uma função, derivada ou a um vetor. A rede G (os matemáticos chamam grafo) é composta de dois subconjuntos $G = (V, L)$, onde V é um conjunto de vértices e L um conjunto de ligações. Para o especialista que for aplicar esta ferramenta, quem serão os vértices, e quem as ligações, fica a critério do modelo. Por hora vamos pensar a rede nesta forma básica: os vértices seriam pontos e as ligações, linhas conectando estes pontos. Apesar das redes serem objetos muito simples, vértices entre si conectados, elas não são vistas no primário: lá estuda-se álgebra, resolução de equações lineares ou geometria cartesiana, conteúdos bem mais sofisticados. Um exemplo visual de rede é a teia de distribuição de eletricidade em um país, onde os vértices são as centrais de distribuição e as conexões são formadas pelas linhas de transmissão. Outra rede bastante em voga é a internet, onde as páginas são os vértices e as ligações se estabelecem através de citações (links) nas páginas. Na Tabela 1 estão ilustradas várias situações que podem ser pensadas usando o conceito de redes. Para cada caso se equaciona: o que são os vértices desta rede, o que são as conexões, e questões que podem ser tratadas com auxílio do modelo.

Tabela 1 - Exemplos de modelos de redes em várias áreas do conhecimento.

Modelo	Vértices	Conexões	Questões de interesse
Rede da energia elétrica [1]	Centrais de distribuição	Linhas elétricas	Estudar resistência da rede frente à pane elétrica
Rede trófica em ecologia [2]	Espécies	Servir de alimento	Caracterizar ecossistemas
Rede das ligações sexuais [3]	Pessoas	Relação sexual	Traçar estratégias de prevenção de epidemias
Rede celular [4]	Metabólitos	Participar de uma mesma reação química	Analisar estabilidade bioquímica da célula
Rede da MPB [5]	Compositores	Intérprete em comum	Investigar o crescimento de grupos sociais
Rede das sílabas do português [6]	Sílabas	Palavra em comum	Conjecturar algorítmos de geração das palavras na língua

Neste artigo trabalharemos com a versão mais simples de modelos de redes, pois apenas estamos interessados em vértices e ligações. Poder-se-ia elaborar um pouco mais o modelo direcionando as ligações, colocado peso nos vértices ou nas ligações. Pelo pouco que acabou de ser exposto o leitor pode perceber o grande potencial de aplicações da rede. A criatividade nesta nova área ainda se encontra longe de ter vislumbrado seus limites.

2. Modelo de rede para análise de textos

A estilometria é o ramo da linguística [7, 8] que dá respostas a perguntas do tipo: qual o escritor que tem a maior riqueza de palavras, Euclides da Cunha ou Rui Barbosa? O procedimento para se responder a esta questão é estatístico, passa pela contagem do número de palavras distintas dos textos escritos pelos autores (normalizado pelo tamanho da obra, ou não). Esta tarefa se tornou relativamente fácil nos dias de hoje com o advento da informática e os bancos de livros virtuais disponíveis na Internet. Enfim, a estilometria essencialmente se baseia em análises estatísticas (rebuscadas ou não) de textos, e o modelo estilométrico apresentado neste artigo não foge a esta regra.

Neste trabalho avaliamos o estilo de um texto sob a ótica da dualidade criatividade versus repetição. Nesta linha poderíamos de uma forma rudimentar quantificar o estilo de um texto contando, por exemplo, de quantas formas diferentes um autor começa suas frases. Um texto estilisticamente pobre seria aquele onde um grande número de frases começa pelos artigos *o* e *a*, pois estas partículas costumam preceder o substantivo que é o principal elemento do sujeito. E no português, em geral, as frases iniciam pelo sujeito. O que faremos neste artigo é usar o conceito de redes para construir uma forma um pouco mais sofisticada de quantificar a variabilidade estilística do que simplesmente fazer o histograma das primeiras palavras de todas as sentenças de um texto.

Vimos na introdução que existe uma grande liberdade para se construirem redes. Nosso objetivo é elaborar uma rede que nos ajude a avaliar a repetição de palavras na mesma posição da frase em uma obra. Esta rede tem a seguinte forma: os vértices são as frases do texto, e as ligações se estabelecem de acordo com a posição relativa das palavras nestas frases. Definimos que dois vértices (frases) estão conectados cada vez que a mesma palavra aparece na mesma posição em cada

uma das frases.

Na Tabela 2 exemplificamos o critério de construção de rede usando oito frases escolhidas ao acaso do romance Brás Cubas de Machado de Assis. Fizemos as seguintes opções no tratamento matemático: não consideramos os pronomes pessoais como palavras distintas do verbo junto ao qual eles se ligam. As diferentes formas declinadas dos verbos e dos nomes são tomadas como palavras diferentes. Ademais, os elementos de pontuação foram desconsiderados nesta análise.

Tabela 2 - Algumas frases V_i tiradas ao acaso do romance Brás Cubas. As palavras foram colocadas em colunas seguindo a aparição na frase.

V_1	Ao	verme	que	primeiro	roeu	as...
V_2	Não	digo	que	se	carpice,	não...
V_3	Sabem,	já	que	morri	numa	sexta-feira...
V_4	Não,	não	me	arrependo	das	vinte...
V_5	Eu	deixei-me	estar	a	contemplá-la.	
V_6	Ao	cabo,	era	um	lindo	garção...
V_7	Mas,	já	que	falei	nos	meus...
V_8	A	que	me	cativou	foi	uma...

A Fig. 1 ilustra a rede linguística formada pelas oito frases do Brás Cubas mostradas na Tabela 2. Nesta rede estão ilustrados os oito vértices representando as frases (V_1, V_2, ... V_8) e suas respectivas ligações. Observe que a forma como os vértices estão distribuidos no plano não tem qualquer importância, apenas as ligações entre os vértices são relevantes. Nesta figura optamos por dispor os vértices sobre uma elipse (esta é umas das possibilidades do nosso software de visualização gráfica [9]). Vale sempre lembrar que uma rede é um objeto que se oferece a infinitas possibilidades de visualização em duas ou três dimensões. Em uma rede os vértices não precisam obedecer a nenhum vínculo métrico no espaço. A única necesidade óbvia que uma representação visual precisa cumprir é: deve haver algum tipo de linha conectando os vértices.

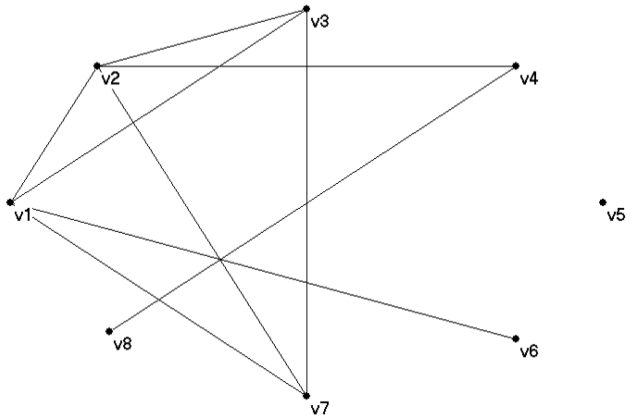

Figura 1 - Rede formada pelas conexões entre as 8 frases da Tabela 2 considerando as ligações entre palavras comuns em qualquer uma das três primeiras posições (caso III).

No exemplo da Fig. 1 consideramos apenas as ligações formadas por palavras que se encontram da primeira até à terceira coluna. Na próxima seção comentaremos com mais vagar quantas colunas são convenientes considerar neste modelo e que tipo de informação se pode extrair da rede assim formada.

3. Análise estatística

Nesta seção analisaremos numericamente a rede que tem por vértices as frases de um texto e cujas ligações se estabelecem pela presença de palavras em comum na mesma posição da frase. Diversamente da maioria dos trabalhos sobre redes não vamos computar o coeficiente de aglomeração (clustering coefficient), nem tampouco a distância média da rede. Usualmente estas grandezas são utilizadas para caracterizar uma rede como complexa, ou não. Neste trabalho nos focaremos nas curvas de estatística de conectividade da rede e destas curvas extrairemos uma ferramenta que nos será útil em estilometria.

A análise numérica foi realizada tomando-se três textos que possuem aproximadamente o mesmo número de frases, ~ 3500. Usamos o Brás Cubas de Machado de Assis, um conjunto diversificado de discursos de Rui Barbosa e um apanhado aleatório de notícias do jornal Tribuna do Norte, um periódico publicado em Natal, RN. Tanto o Brás Cubas, como os escritos de Rui Barbosa, foram baixados do *site* da Biblioteca Virtual da USP [10].

A distribuição estatística que utilizaremos está mostrada na Fig. 2. Na abscissa se encontram os vértices, i, em ordem decrescente de conectividade, e no eixo vertical seus respectivos números de conexões, k_i. A Fig. 2 (a) corresponde aos dados do texto de Machado de Assis enquanto 2 (b) aos da Tribuna do

Norte. Em ambas as figuras são apresentadas quatro curvas. A primera é a menos suave (caso I) e se refere à situação em que a rede é construída se tomando em conta apenas as ligações relativas às palavras que se encontram na primeira posição da frase. A segunda curva (caso II) corresponde à rede construída a partir de ligações feitas se tomando em conta a primeira e a segunda posições da oração. A terceira e quarta curvas (casos III e IV) são feitas se usando, respectivamente, as três e quatro primeiras posições na frase. Não apresentamos as curvas relativas ao texto de Rui Barbosa por serem muito semelhantes às de Machado de Assis; assim concentraremos nossa atenção inicialmente em observar as diferenças entre as figuras relativas aos textos de Machado de Assis e da Tribuna do Norte.

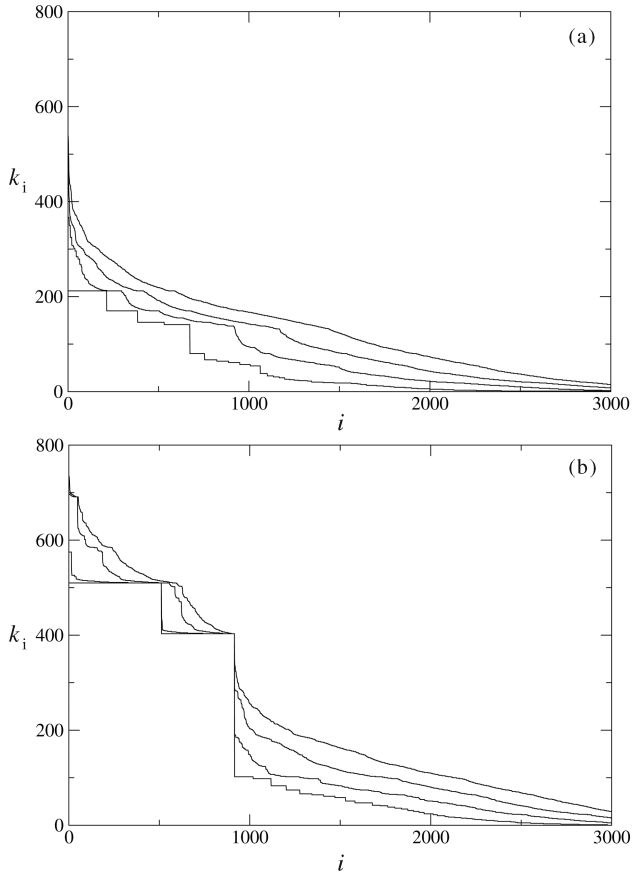

Figura 2 - No eixo horizontal os vértices em ordem decrescente de conectividade, i, no eixo vertical os respectivos número de conexões, k_i. As quatro curvas correspondendo aos casos I, II, III e IV. Em (a) são mostradas as curvas do Brás Cubas de Machado de Assis e em (b) as curvas referentes a um apanhado de textos do jornal Tribuna do Norte. Os textos têm aproximadamente o mesmo número de palavras.

Observando a Fig. 2 se nota que, tanto em (a) como em (b), à medida que passamos do caso I em direção ao IV o número total de ligações na rede aumenta e as curvas vão se tornando cada vez mais suaves. Um aspecto que salta aos olhos na comparação entre as Figs. 2 (a) e (b) é a presença mais acentuada de degraus em (b). Vamos nos concentrar incialmente no caso I, que realiza a estatística das frases conectadas entre si pelo fato de começarem pela mesma palavra. O primero degrau do caso I corresponde a todas as frases que iniciam pelo artigo definido *o*, pois trata-se da forma mais comum de se começarem frases em português. O segundo degrau refere-se às frases iniciadas em *a*, e assim por diante. Explica-se então por que a figura machadiana, 2 (a), é, como um todo, mais suave do que a da Tribuna do Norte. Machado de Assis principia suas orações de uma forma muito mais variada e criativa do que os apressados jornalistas da Tribuna.

Uma reflexão sobre a Fig. 2 indica que um texto pobre apresenta muito mais descontinuidades do que um texto rico em estilo. Um autor de estilo rico usa um maior número de palavras e as joga com mais liberdade ao longo das posições da frase. A relativa pobreza de estilo, que nos propomos a quantificar, se evidencia na forma jornalística de se construir frases: "O assaltante atirou três vezes...", "O policial reagiu prontamente a...", ou "Um assaltante mulato pulou o muro da...". Com o intuito de quantificar os saltos nas curvas da Fig. 2 calcularemos a dispersão, D, de k_i definida como segue:

$$D = \frac{\sum_i^N (k_{i+1} - k_i)^2}{N}, \qquad (1)$$

onde N é o número total de frases. A grandeza D, pelo fato de estar normalizada, crescerá quanto mais pronunciadas forem as descontinuidades das curvas estudadas. Na Tabela 3 mostramos o cálculo de D para os três textos analisados. Duas observações sobre esta tabela. A primeira é que, de fato, Rui Barbosa e Machado de Assis apresentam, como era de se esperar, valores de D muito menores do que a Tribuna nos casos de I a III. A segunda observação diz respeito à caracterização de aleatoriedade usando estatística, os textos apresentados não mostram diferença em D para o caso IV. Interpretamos este fato pelo aleatoriedade das palavras presentes a partir da quarta posição na sentença.

Tabela 3 - A dispersão, D, da curva da soma cumulativa para os três textos estudados. Avaliamos os resultados para redes construídas usando primeira, segunda, terceira e quarta colunas, isto é, casos I a IV.

	I	II	III	IV
Machado de Assis	2,5	0,76	1,7	1,1
Rui Barbosa	2,6	0,67	1,1	0,46
Tribuna do Norte	35	17	5,6	1,6

4. Considerações finais

Esperamos neste artigo ter apresentado ao leitor uma rede singular formada com auxílio das posições relativas das palavras dentro das frases de um texto. Usamos esta rede como uma ferramenta de análise do estilo de

um texto. Através de alguns exemplos mostramos que este modelo pode ser útil para caracterizar sua pobreza ou riqueza de estilo. Uma palavra de cautela: os autores não reivindicam que este método deva ser usado para julgar autores sem contar com o apoio de outras interpretações e outros olhares com diferentes visões de mundo. Uma obra de arte (e um texto pode pertencer a este grupo) é um objeto aberto a muitas interpretações e leituras, e por certo não nos agradaria ver este método de estilometria sendo usado isoladamente na tarefa de julgar o caráter literário de textos (por mais que algum de nós se sinta polemicamente tentado a aplicá-lo aos imortais de nossa academia).

Sobretudo esperamos que este trabalho sobre redes tenha sido de serventia a todos aqueles que queiram conhecer um pouco mais sobre recentes desenvolvimentos de métodos da Física. Não nos referimos ao método já aplicado e consumado sobre um objeto antigo, mas a ele vivo, sendo utilizado em um novo campo, criando novas perguntas e inspirando novas aplicações. Nossa tarefa terá sido bem sucedida se tivermos mostrado ao leitor o quão diversas são e, ainda mais, podem vir a ser as aplicações da teoria de redes.

Não poderíamos finalizar o trabalho sem comentar algo sobre a distribuição de conectividade, $p(k)$, da rede estudada. $p(k)$ é o histograma da conectividade: esta grandeza computa quantos vértices existem com uma dada conectividade k. Ora, os primeiros trabalhos sobre redes [11] já utilizavam $p(k)$ para estabelecer uma classificação geral das redes. Se $p(k)$ assumir um valor constante temos uma rede regular como as redes cristalinas. Quando $p(k)$ apresentar uma distribuição gaussiana, a rede se denomina aleatória e neste caso a probabilidade de quaisquer dois vértices se conectarem entre si é randômica. O caso mais interessante para os olhos do século XXI, contudo, é quando $p(k)$ assumir a forma de uma distribuição do tipo lei de potência. Então a rede se chamaria complexa, ou tipo pequeno mundo. Escrevemos a frase anterior no futuro do pretérito, pois existe uma certa controvérsia sobre este ponto. Alguns autores preferem que uma rede seja chamada de complexa quando possuir um coeficiente de aglomeramento grande se comparado com uma rede aleatória correspondente.

Em nosso trabalho não usamos diretamente $p(k)$, mas a soma cumulativa (Fig. 2). Para obtermos $p(k)$ basta tomarmos a derivada, ou a diferença no caso discreto, da soma cumulativa. Vale notar que na classificação de redes do parágrafo anterior, $p(k)$ (para redes suficientemente grandes) tende sempre a uma curva suave. Nosso modelo, pelo contrário, tira suas melhores conclusões, justamente, das descontinuidades na soma cumulativa. Tanto quanto saibam os autores esta é a primeira aplicação da teoria de redes onde o caráter não suave na distribuição das conexões é relevante na análise dos resultados.

Finalizamos este trabalho retornando ao primeiro parágrafo que começava afirmando ser este artigo destinado a dois públicos bem diversos: os curiosos por teoria de redes e os interessados em estilometria. Para não cometer uma injustiça deveria ainda computar um terceiro grupo: os amantes de Machado de Assis - aqueles que se deixam deleitar com frases machadianas encontradas, até mesmo, numa tabela perdida em um artigo da Revista Brasileira de Ensino de Física.

Agradecimentos

Os autores agradecem à professora Suani Pinho pelo estímulo em escrever este trabalho e ao professor Liacir Lucena pelo apoio nos momentos críticos. Agradecemos também pela utilização do software de processamento de redes Pajek [9].

Referências

[1] L.A.N. Amaral, A. Scalam M. Barthélémy and H.E. Stanley, Proc. Natl. Acad. Sci. USA **97**, 11149 (2000).

[2] J.M. Montoya and R.V. Solé, Journal of Theoretical Biology **214**, 405 (2002).

[3] F. Liljeros, C.R. Edling, L.A.N. Amaral, H.E. Stanley and Y. Aberg, Nature **411**, 907 (2001).

[4] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai and A.-L Barabási, Nature **407**, 651 (2000).

[5] D. de Lima e Silva, M. Medeiros Soares, M.V.C. Henriques, M.T. Schivani Alves, S.G. de Aguiar, T.P. de Carvalho, G. Corso and L.S. Lucena, Physica A **311**, 590 (2004).

[6] M. Medeiros Soares, G. Corso and L.S. Lucena, aceito em Physica A **355**, 678 (2005).

[7] Jean Dubois, *Dicionário de Linguística* (Cultrix, São Paulo, 1973).

[8] J. Lyons, *Language and Linguistics* (Cambridge Univ. Press, Cambridge, 1990).

[9] http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

[10] http://www.bibvirt.futuro.usp.br/index.html

[11] D.J. Watts and S.H. Strogatz, Nature, **393**, 440 (1998).