Uma Investigação acerca da Conectividade da Web Brasileira

Cristina D. Murta[1], **Valter R. Lima Jr.**[1], **Adriano C. M. Pereira**[2]

[1] Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
30.510-000 – Belo Horizonte – MG – Brasil

[2]Universidade Federal de Minas Gerais (UFMG)
31.270-010 – Belo Horizonte – MG – Brasil

Resumo. *Este artigo apresenta uma análise da conectividade e da estrutura topológica da Web brasileira e da Web governamental brasileira, realizada a partir de duas coletas recentes feitas por robôs. As coletas realizadas alcançaram cerca de 7% dos domínios registrados oficialmente no país. Os dados coletados foram filtrados e transformados em grafos, que foram analisados de acordo com várias métricas. Os resultados indicam que a Web brasileira contém um componente central fortemente conectado que engloba cerca de apenas 11% dos seus vértices. Há uma grande disparidade na densidade de conexões internas aos sítios e conexões entre sítios. A análise conjunta dos resultados indica que a conectividade da Web brasileira é baixa.*

Abstract. *This paper presents an analysis of the connectivity and the topological structure of the Brazilian Web and the Brazilian Government's official Web, made from two datasets recently collected by Web crawlers. The samples taken encompass about 7% of the Web domains officially registered in the country. The collected data were filtered and transformed into graphs, which were analyzed according to various metrics. The results indicate that the Brazilian Web contains a strongly connected component which includes only 11% of its vertices. There is a wide disparity in the density of connections internal to Web sites and connections between Web sites. The analysis of the results show that the connectivity of the Brazilian Web is low.*

1. Introdução

A presente revolução digital está alterando profundamente as maneiras de viver das sociedades, em particular quanto ao fluxo e a troca de informação. A WWW ou Web é, por excelência, a aplicação da Internet mais relevante nesse contexto. Imaginada por Vannevar Bush em 1945 [Bush and Wang 1945] como uma memória coletiva e concebida por Tim Berners-Lee [Wikipedia 2004] por volta de 1990, tendo como base o conceito de hipertexto, a Web é atualmente uma ferramenta utilizada intensivamente por uma fração importante da população mundial. Por ser um meio pouco controlado, qualquer indivíduo ou organização pode criar sítios sem qualquer restrição em relação à quantidade de páginas ou de *links* [Barabási et al. 2000a]. Este fato ocasionou seu crescimento irregular, transformando a Web em uma rede complexa de grande escala e grande dinamicidade. Devido a essas características, a análise de suas propriedades requer estudos contínuos, que incluem coleta e processamento de quantidades gigantescas de dados para extração de informação.

A conectividade de uma rede tem papel importante na disseminação de informação por meio dessa rede. A troca e o fluxo de informação eficientes requerem

redes bastante conectadas. No caso da Web, o caminhamento na rede depende fundamentalmente de sua conectividade. O presente artigo busca analisar a seguinte questão: quão conectada é a Web brasileira? Para responder a essa questão, foram realizadas e analisadas duas coletas da Web brasileira. A primeira coleta teve como alvo páginas identificadas pelo domínio `.br`. A segunda coleta foi feita especificamente para a Web governamental, identificada pelo domínio `.gov.br`. Os dados coletados foram transformados em grafos, cujas métricas foram analisadas.

A Web é constituída por um conjunto de documentos denominados páginas que são escritas de acordo com a linguagem de marcação HTML. Essas páginas possuem conexões virtuais para outras páginas ou recursos, denominadas *hyperlinks* ou apenas *links*. Assim, a Web pode ser modelada por um grafo dirigido, em que os vértices são páginas HTML e as arestas dirigidas são *links* que apontam de uma página para outra. Cada página tem um endereço dado por uma URL (*Uniform Resource Locator*).

Os primeiros estudos acerca da caracterização e da estrutura da Web surgiram entre 1999 e 2000 [Barabási and Albert 1999, Broder et al. 2000, Kumar et al. 2000, Barabási et al. 2000a]. Várias características da Web foram apontadas nesses trabalhos. Por exemplo, foram identificadas leis de potência na distribuição dos graus dos vértices, que são resultado do processo de conexão preferencial. As distâncias entre os vértices são pequenas e se enquadram no fenômeno *small-world*. A rede apresenta um único componente fortemente conectado, muito grande, e vários outros pequenos. Além disso, o modelo gravata borboleta foi proposto para representar sua estrutura topológica [Broder et al. 2000]. A Web brasileira foi objeto de dois estudos importantes [Veloso et al. 2000, Modesto et al. 2005], cujo enfoque foi caracterizar o conteúdo das páginas coletadas, bem como avaliar sua estrutura.

Nesse artigo apresentamos uma investigação acerca da conectividade e da estrutura topológica da Web Brasileira, a partir de coletas recentes de dados. O objetivo é avaliar como a Web brasileira está conectada e se a rede está conectada de maneira que seja possível percorrê-la sem a ajuda de máquinas de busca. Os dados coletados foram transformados em grafos. Vários algoritmos de análise de grafos foram utilizados para obtermos diversas métricas de conectividade e de caracterização dos grafos formados. A estrutura da Web brasileira atual foi também avaliada de acordo com o modelo gravata [Broder et al. 2000]. Outro objetivo é verificar como a Web governamental brasileira se compara à Web brasileira como um todo.

Os resultados indicam que a Web brasileira e a Web governamental brasileira apresentam características qualitativas similares às apontadas nos estudos anteriores. Porém, a análise numérica e conjunta dos resultados revela uma rede pouco conectada. Em particular, há uma enorme disparidade entre a densidade das conexões internas aos sítios e as conexões entre sítios. A relevância da análise é dada pela expressiva fração da Web coletada, que alcança mais de 7% de todos os domínios oficialmente registrados no país.

Este artigo está organizado em cinco seções. A próxima seção apresenta os trabalhos relacionados ao tema e discute a inserção desse trabalho no contexto do conhecimento da área. A Seção 3 apresenta a metodologia de pesquisa utilizada para se obter os resultados, que são apresentados na Seção 4. A Seção 5 apresenta as conclusões obtidas no estudo, bem como as indicações para trabalhos futuros.

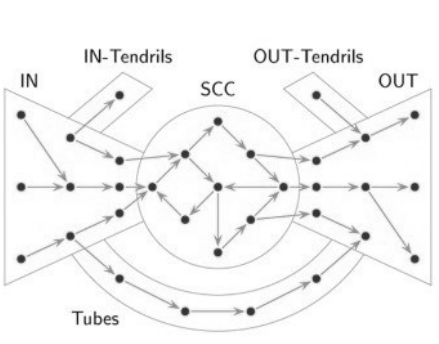


Figura 1. Componentes de um grafo da Web [Vitali et al. 2011].

2. Conceitos e Trabalhos Relacionados

A estrutura da Web pode ser representada por um grafo dirigido, em que os vértices são páginas HTML e os arcos são *links* apontando de uma página para outra [Boccaletti et al. 2006]. Outra maneira de modelar esse grafo é agrupar as páginas por sítios, formando um grafo em que os vértices são sítios e arestas são conexões entre sítios. Após a modelagem, a análise é feita por algoritmos que fazem medições em grafos, seguindo o procedimento de análise de redes complexas [Boccaletti et al. 2006].

Uma característica marcante da estrutura da Web é o fato de que a distribuição dos graus (de entrada e de saída) dos vértices segue uma lei de potência [Barabási et al. 2000a, Barabási et al. 2000b, Chayes 2013], dada pela equação $P(k) \sim k^{-\gamma}$ em que k é o grau e γ é uma constante que varia entre 2 e 3 [Boccaletti et al. 2006, Donato et al. 2007, Barabási and Albert 1999]. Outra característica importante da Web é sua dinamicidade, decorrente do fato de que novas páginas são adicionadas ou excluídas e *links* são criados e removidos continuamente. Estudos indicam que novas páginas tendem a se conectar a páginas populares, seguindo um processo conhecido como conexão preferencial, o que ocasiona grande variabilidade nos graus dos vértices [Barabási et al. 2000a].

O estudo dos componentes do grafo da Web revela o que alguns autores chamam de "cyber-comunidades" [Broder et al. 2000]. A Figura 1 apresenta um grafo desenhado no modelo conhecido como "gravata borboleta". Há um componente fortemente conectado (SCC) desenhado na posição central do grafo. O conjunto IN é um componente fracamente conectado composto por vértices a partir dos quais é possível alcançar os vértices do SCC. O componente fracamente conectado OUT consiste dos vértices que podem ser alcançados a partir de qualquer vértice contido em SCC. O conjunto Tubes conecta vértices de IN diretamente aos de OUT. O conjunto Tendrils inclui vértices que se ligam a vértices desconectados do conjunto SCC. Por fim, há um conjunto, não representado na Figura, de vértices e de pequenos componentes sem conexão com o grafo principal [Donato et al. 2007, Broder et al. 2000].

Em um estudo pioneiro, um grafo de cerca com 200 milhões de vértices e 1,5 bilhão de arcos, obtido em coleta realizada de maio a outubro de 1999, foi analisado segundo o modelo gravata [Broder et al. 2000]. A análise do grafo indicou que o componente SCC incluía cerca de 56 milhões de vértices (o maior componente encontrado). No componente IN residiam cerca de 44 milhões de vértices e, no componente OUT, outros 44 milhões de vértices. Os vértices restantes compunham os demais conjuntos.

Uma análise posterior, de um grafo com cerca de 200 milhões de vértices e 1,4 bilhão de arcos, obteve resultados distintos [Donato et al. 2007]. Nesse caso, o componente SCC representou cerca de 33% do grafo e o componente IN incluiu cerca de 11%. O maior componente contido no grafo foi o OUT, com cerca de 39% do grafo. Os componentes Tubes e Tendrils representaram 13% e o restante compôs o grupo desconectado.

Uma métrica essencial no estudo da topologia da Web é a distância média entre pares de vértices. Esta métrica consiste em calcular a menor distância entre todos os pares de vértices do grafo e, posteriormente, obter a média das menores distâncias encontradas. Estudos prévios em grafos da Web encontraram valores para a distância média de 18,59 [Barabási et al. 2000a] e 19 [Broder et al. 2000].

A Web brasileira foi objeto de estudo em dois trabalhos publicados no SEMISH. O trabalho pioneiro de [Veloso et al. 2000] analisou o conteúdo e a estrutura da Web brasileira no ano 2000. Dentre os resultados apresentados estão os tipos de documentos encontrados (HTML, PDF, MS Word, etc.) e também suas quantidades. O estudo se estende para identificar os idiomas mais comuns na Web brasileira. Ainda neste trabalho são exibidos os tipos e quantidades de Domínios de Primeiro Nível (DPN) encontrados na coleta. A quantidade média de *links* por página foi calculada em 6,74, o que corresponde ao grau médio dos vértices. Por fim, os autores estimaram o tamanho da Web brasileira na época, em relação à quantidade de páginas, chegando a um valor de quase dezoito milhões.

O segundo estudo encontrado sobre a Web brasileira [Modesto et al. 2005] apresenta um novo retrato dessa rede complexa, revelando seu conteúdo e estrutura, a partir de dados de uma coleta mais atualizada, de março de 2005. As análises realizadas nesse trabalho foram praticamente as mesmas do trabalho anterior [Veloso et al. 2000], com o objetivo de descobrir as mudanças na Web brasileira no período de cinco anos. O grafo obtido foi modelado de acordo com o modelo gravata já descrito, e cada componente obteve a seguinte fração dos vértices do grafo: componente SCC com 25,27%; componente IN com 12,95%; componente OUT com 45,33%; componente Tendrils com 3,87%; componente Tubes com 0,23% e o conjunto desconectado com 12,35% do grafo.

A Web portuguesa, definida por páginas no domínio `.pt`, também foi objeto de um estudo similar [Gomes and Silva 2003]. A coleta resultou em cerca de 4 milhões de páginas de 131 mil sítios. Os autores constataram que 85% da Web portuguesa está contida no domínio `.pt` e que os 15% restantes se tratam de sítios portugueses contidos em domínios como `.com`, `.net`, `.org` e `.tv`.

Curiosamente, não se observa na literatura muitos artigos com caracterizações de redes Web, possivelmente pelo fato da constatação de seu gigantismo [Google 2008], bem como pela complexidade dos processamentos e análises. No entanto, o aspecto dinâmico da Web, em mudança contínua, bem como sua importância para a sociedade atual, requerem esse tipo de estudo para seu melhor entendimento.

3. Coleta de Dados e Metodologia

A presente análise da Web brasileira baseia-se em coletas feitas pelo Web *crawler* Heritrix [Jack and Binns 2012]. Um *crawler* é um software que trabalha como um robô, percorrendo a Web a partir de um conjunto de sítios iniciais chamados sementes. O Heritrix faz buscas em profundidade, seguindo as URLs descobertas, registrando e armazenando

Tabela 1. Processo de extração dos sítios a partir das URLs.

Antes	Depois
http://www.cefetmg.br/site/instituicao/	www.cefetmg.br
http://tvmissoes.com.br/fotos/displayimage.php	tvmissoes.com.br
http://maps.google.com.br/about/	maps.google.com.br
http://www.uol.com.br/museus/picasso/	www.uol.com.br
http://www.baixaki.com.br/jogos/	www.baixaki.com.br
http://www.yahoo.com.br/games	www.yahoo.com.br
http://www.youtube.com.br/NEToficial	www.youtube.com.br

os dados encontrados em um arquivo. Dentre as informações registradas está o caminho percorrido pelo *crawler* na Web, URL a URL. O resultado da coleta é um arquivo que representa uma lista de arcos contendo uma URL como vértice de origem e outra URL como vértice de destino, dentre outras informações. Devido ao gigantesco volume de dados da Web, o processo de coleta não termina com a coleta de toda a Web, e sim quando os recursos computacionais dedicados à coleta são exauridos. Assim, a coleta é finalizada.

Os endereços Web são organizados de forma hierárquica. Nosso foco é a Web brasileira, cujo domínio principal tem a terminação `.br`. Seguindo o string da URL, temos o domínio de primeiro nível, que pode ser exemplificado, no caso do Brasil, pelas seguintes terminações: `com.br`, `gov.br`, `net.br`, dentre outras. A lista oficial de domínios de primeiro nível (DPN) válidos no país é definida pelo Comitê Gestor da Internet no Brasil[1]. Há dezenas de opções em várias categorias. Seguindo a hierarquia, temos os nomes registrados pelos usuários, que estão sempre incluídos em um dos DPN. Por exemplo, `uol.com.br` e `mg.gov.br` são domínios registrados pelos usuários. No próximo nível, temos os subdomínios, por exemplo, `fotoblog.uol.com.br` e `fazenda.mg.gov.br`, definidos no âmbito do usuário.

Para realizar esse trabalho foram feitas duas coletas da Web brasileira. A primeira coleta foi realizada no período de 9/12/2012 a 14/01/2013 e incluiu todas as páginas alcançadas dentro do domínio principal `.br`. A segunda coleta foi realizada durante o mês de agosto de 2013 e teve seu foco na Web governamental brasileira, definida pela terminação `.gov.br`.

Uma vez realizadas as coletas, o próximo passo foi modelar os grafos. No processo de modelagem foram incluídas apenas as URLs que retornaram códigos de status HTTP nas classes 2xx e 3xx, indicando sucesso no acesso à página ou redirecionamento. Esse processo produziu um arquivo em que cada linha continha um arco dirigido de uma URL para outra URL, sendo ambas as URLs válidas. Após esse processo, os arquivos de dados somaram, respectivamente, 75 Gbytes e 60 Gbytes.

O alvo de nossa análise é um grafo em que cada vértice é um sítio ou subsítio, não o de grafo de páginas. Assim, a etapa seguinte foi a construção do grafo de sítios e subsítios, a partir do grafo de páginas. Para isso extraímos de cada URL o string contido entre as duas barras após o protocolo de acesso e a primeira barra que simboliza o primeiro diretório do sítio. Exemplos deste processo podem ser visualizados na Tabela 1.

No arquivo original obtido pelo Heritrix não há URLs repetidas, o que é devido ao processo de caminhamento realizado pelo *crawler*. No entanto, o processo de extração

[1] www.cgi.br

Tabela 2. Formação de laços no processo de extração dos sítios.

Estado	Vértice de Origem	Vértice de Destino
Antes	http://tvmissoes.com.br/fotos/	http://tvmissoes.com.br/robots.txt
Depois	tvmissoes.com.br	tvmissoes.com.br
Antes	http://www.google.com.br/intl/af/	http://www.google.com.br/intl/af/policies/
Depois	www.google.com.br	www.google.com.br

Tabela 3. Formação de arcos repetidos no processo de extração dos sítios.

Estado	Vértice de Origem	Vértice de Destino
Antes	http://www.uol.com.br/	http://esporte.uol.com.br/futebol/clubes/corinthians/
	http://www.uol.com.br/	http://esporte.uol.com.br/futebol/clubes/boa-esporte-clube/
Depois	www.uol.com.br	esporte.uol.com.br
	www.uol.com.br	esporte.uol.com.br

dos sítios a partir das strings das URLs produz arcos repetidos e laços, que são arcos que têm origem e destino no mesmo sítio. A Tabela 2 apresenta um exemplo de como um laço é formado nesse processo. Observamos que, inicialmente, duas URLs distintas formam um arco dirigido. No entanto, essas URLs pertencem ao mesmo sítio, o que gera um laço, que é uma conexão interna ao próprio sítio.

O processo de extração dos sítios gera também arestas repetidas, que não ocorriam no grafo de páginas. A Tabela 3 mostra como são geradas arestas repetidas no grafo de sítios. Arcos distintos obtidos na coleta geram arcos idênticos (mesma origem e destino) após o processo de redução dos strings das URLs.

Os laços e arcos repetidos foram contados e posteriormente retirados do grafo. Ao fim do processo de modelagem, os grafos estão armazenados em arquivo na forma de lista de arcos, que representam conexões entre sítios ou subsítios. Todos os arcos são distintos e não há laços. As dimensões dos grafos obtidos são apresentadas na Seção 4.

Para a análise da conectividade, várias métricas de análise de redes complexas foram avaliadas, dentre elas o tamanho e a densidade da rede, estatísticas dos graus e das distâncias, diâmetro da rede, número e tamanhos dos componentes fortemente conectados e análise topológica da rede. Essas métricas foram calculadas de tanto por programas implementados por nós [2], quanto usando o software Pajek [Pajek 2013], para conferência. Os resultados são apresentados na próxima seção.

4. Resultados

Essa seção apresenta os resultados da análise dos grafos gerados a partir das coletas.

4.1. Dimensão dos Grafos

Conforme exposto na seção anterior, a partir das coletas obtivemos os grafos de páginas e a seguir fizemos a redução do string da URL para obter as conexões (*links*) entre sítios e subsítios, que são os novos vértices dos grafos. A Tabela 4 apresenta a quantidade inicial de arcos nos dois grafos `.br` e `.gov.br`, bem como o número de laços e de arcos repetidos encontrados no processo de extração dos sítios. Esta Tabela permite análises importantes acerca da conectividade da Web brasileira.

[2]https://github.com/valtincomp/implementacoes-semish

Tabela 4. Dimensões dos grafos coletados

Medida	Grafo .br	Grafo .gov.br
Arcos	260.927.712	138.366.042
Laços	253.395.811	137.180.945
Arcos repetidos	6.387.326	1.138.901
Arcos únicos	1.140.004	45.304
Vértices	519.917	19.523

No caso do grafo `.br`, temos inicialmente 260.927.712 arcos. Porém, a ampla maioria deles (97,11%) é de laços, ou seja, conexões intra-sítios (de uma URL do sítio para outra URL do mesmo sítio). Somente 7.527.330, resultado da soma de arcos únicos e arcos repetidos, são *links* entre sítios distintos, o que corresponde a 2,89% dos *links*. No caso do grafo `.gov.br`, o resultado é ainda mais extremo, com 99,14% de conexões intra-sítios e 0,86% são ligações entre sítios.

Nesse cenário, é interessante analisar a conectividade interna aos sítios em comparação com a conectividade entre sítios. No caso do grafo `.br`, o número de laços dividido pela soma do número de arcos únicos mais arcos repetidos é igual a 33,66. Ou seja, para cada arco entre sítios distintos há cerca de 33 conexões internas aos sítios. Além disso, o número médio de laços por vértice é 487,38. Há, portanto, muito mais conexões internas aos próprios sítios do que conexões entre sítios diferentes. Em outras palavras, sítios e subsítios são muito mais conectados internamente do que entre si.

O número de vértices em cada grafo é também apresentado na Tabela 4. Em termos do número de vértices e de arestas, o grafo da Web governamental brasileira é pequeno em relação ao grafo da Web brasileira. Ele corresponde a 3,8% do número de vértices e cerca de 4% do número de arestas do grafo maior.

4.2. Estatísticas dos Graus e das Distâncias

A análise dos grafos é iniciada pela análise dos graus de entrada e de saída. A Tabela 5 apresenta as estatísticas dos graus, a saber, menor e maior graus, grau médio e mediana, e também o coeficiente de variação (COV) para ambos os grafos. Os graus médios de entrada e de saída são iguais uma vez a soma dos graus de entrada é igual à soma dos graus de saída. A mediana é próxima à média. O menor grau é zero em todos os casos. Quanto ao maior valor, os graus de saída alcançam valores muito superiores aos graus de entrada, o que eleva o coeficiente de variação (definido pela razão entre o desvio padrão e a média). Enquanto no grafo da Web governamental os maiores graus são da ordem de centenas, na Web brasileira os maiores graus são da ordem de milhares. É importante notar que o grau médio, com valores pouco acima de dois, é muito baixo, pois é apenas uma unidade acima do mínimo para conectar um vértice ao grafo.

Tabela 5. Estatísticas dos graus para ambos os grafos

Grafo	Métrica	Menor	Maior	Média	Mediana	COV
`.br`	Grau de entrada	0	2.977	2,19	1	4,65
`.br`	Grau de saída	0	76.351	2,19	0	55,96
`.gov.br`	Grau de entrada	0	343	2,32	1	3,39
`.gov.br`	Grau de saída	0	760	2,32	0	5,16

A Tabela 6 apresenta os dez vértices com os maiores graus de entrada e

de saída para o grafo `.br`. Observa-se na primeira coluna dois subsítios do sítio `www.softonic.com.br`. Na segunda coluna, que contém os vértices com os maiores graus de saída, foram identificados quatro sítios especializados em realizar algum tipo de busca na Web.

Tabela 6. Os 10 vértices com os maiores graus do grafo .br

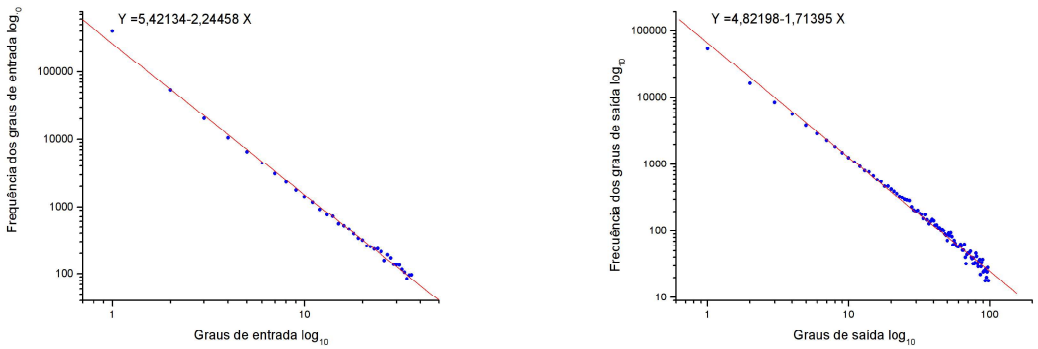
Sítio	Grau IN	Sítio	Grau OUT
fotoblog.uol.com.br	2.977	www.lyrics.com.br	76.351
onsoftware.softonic.com.br	1.644	www.blogorama.com.br	34.591
selos.climatempo.com.br	1.536	www.letrasdemusicas.com.br	22.166
cifraclub.terra.com.br	1.495	fotoblog.uol.com.br	9.706
dplus.softonic.com.br	1.474	www.olx.com.br	4.469
contador.s12.com.br	1.013	www.guiademidia.com.br	4.413
maps.google.com.br	928	aonde.com.br	3.988
www.band.com.br	803	www.guis.com.br	3.134
www.orkut.com.br	754	www.softonic.com.br	3.119
www.vagas.com.br	732	www.somanuncios.com.br	2.834

A Tabela 7 apresenta os dez vértices com os maiores graus de entrada e de saída para o grafo `.gov.br`. O sítio com o maior grau de entrada é o da Imprensa Nacional, que contém e divulga as informações oficiais do país. De forma geral, os vértices com os maiores graus de entrada são órgãos do governo federal bem como do poder legislativo. Por outro lado, dentre os vértices com os maiores graus de saída, há órgãos dos governos estaduais do PR, SP, MG e SC.

Tabela 7. Os 10 vértices com os maiores graus do grafo .gov.br.

Sítio	Grau IN	Sítio	Grau OUT
www.in.gov.br	343	www.nre.seed.pr.gov.br	760
portal.mec.gov.br	287	www.bibliotecavirtual.sp.gov.br	716
www.ibge.gov.br	255	www.cidadao.pr.gov.br	535
www2.camara.gov.br	245	www.comunidade.diaadia.pr.gov.br	381
www.senado.gov.br	242	dominios.governoeletronico.gov.br	286
www.planalto.gov.br	239	portaldoprofessor.mec.gov.br	230
portal.saude.gov.br	237	www.turismo.mg.gov.br	223
www.camara.gov.br	213	www.santur.sc.gov.br	221
www.portaldatransparencia.gov.br	173	www.almg.gov.br	212
www.mj.gov.br	166	www.servicos.gov.br	205

Os gráficos da Figura 2 apresentam, para o grafo `.br`, a frequência dos graus de entrada e de saída em escala logarítmica. Nesse tipo de gráfico, o decaimento da frequência dos graus em linha reta, conforme mostrado, indica que os graus podem ser modelados por uma lei de potência, o que é uma evidência do modelo conhecido como livre de escala, gerado a partir de conexões feitas preferencialmente aos vértices com maiores graus [Barabási and Albert 1999]. O ajuste dos pontos resultou que, para os graus de entrada, foi encontrado o exponente 2,24, enquanto para os graus de saída foi identificado o exponente 1,71. Os mesmos gráficos foram feitos para o grafo `.gov.br`, e foram encontrados, para a distribuição dos graus de entrada, o expoente 2,28, e, para a distribuição dos graus de saída, o expoente 1,72. Por questão de falta de espaço esses gráficos não são apresentados. Observamos que a lei de potência está presente em ambos os grafos, tanto nos graus de entrada quanto nos graus de saída, e os valores numéricos dos expoentes são próximos.



(a) Graus de entrada

(b) Graus de saída

Figura 2. Leis de potência nos graus do grafo .br.

Uma medida fundamental de conectividade em grafos é dada pelas estatísticas da métrica distância entre pares de vértices. A distância, nesses grafos, indica o número de *links* que devem ser seguidos para alcançar um vértice a partir de outro, considerando sempre o menor caminho entre eles. Para isso, foram calculadas as distâncias mínimas entre todos os pares de vértices, em cada grafo. A Tabela 8 apresenta dados estatísticos das distâncias para os dois grafos. Os dados indicam que as distâncias do grafo `.br` são maiores do que as do grafo da Web governamental (exceto o menor valor), refletindo sua maior dimensão. A distância média em ambos os grafos é baixa, o que é uma evidência do fenômeno *small-world*. Porém, se a rede fosse mais conectada, a distância média e a maior distância seriam ainda menores. As maiores distâncias indicam os diâmetros dos grafos analisados, ou seja, os maiores caminhos entre dois vértices, em cada grafo. Os valores próximos da média e da mediana e o baixo coeficiente de variação indicam uma distribuição das distâncias aproximadamente simétrica.

Tabela 8. Dados estatísticos das distâncias

Grafo	Menor	Maior	Média	Mediana	COV
`.br`	1	26	7,66	7	1,24
`.gov.br`	1	19	5,91	6	1,16

4.3. Análise da Topologia da Web Brasileira

Outra medida da conectividade dos grafos é dada pelo número e tamanho dos componentes fortemente conectados (CFC), dados que são apresentados na Tabela 9. No caso do grafo `.br` foram descobertos 1.087 CFCs, sendo que o maior CFC descoberto contém 57.063 vértices. Vinte e dois CFCs contém dezenas de vértices e 1.064 CFCs são pequenos, com um número de vértices na casa das unidades. Há portanto, um único e grande componente fortemente conectado. O segundo maior CFC contém apenas 70 vértices. Situação similar ocorre com o grafo `.gov.br`, que também contém um componente conectado com 5.234 vértices. O segundo maior CFC contém apenas 16 vértices. Os demais CFCs são ainda menores.

Tabela 9. Número e tamanhos dos componentes fortemente conectados

Medida	Grafo .br	Grafo .gov.br	Explicação
Maior CFC	57.063	5.234	Número de vértices no maior CFC
CFCs médios	22	1	Número de CFCs com dezenas de vértices
CFCs pequenos	1.064	198	Número de CFCs com menos de 10 vértices

Tabela 10. Dimensões do modelo gravata nos grafos

Componente	Grafo .br		Grafo .gov.br	
	No. Vértices	Percentual	No. Vértices	Percentual
SCC	57.063	11,00	5.234	26,81
IN	12.829	2,47	623	3,19
OUT	290.646	56,05	12.776	65,44
Tubes	4.943	0,95	63	0,32
Tendrils	152.250	29,36	761	3,90
Desconectados	835	0,17	66	0,34

Os resultados desta métrica revelam outra característica do grafo da Web brasileira em consonância com relatos da literatura. Conforme estudos anteriores, o grafo da Web contém um grande componente fortemente conectado e vários outros componentes com pequeno número de vértices [Broder et al. 2000, Modesto et al. 2005, Donato et al. 2007].

A identificação da estrutura topológica dos grafos também foi feita. Utilizando o software Pajek, medimos as dimensões do grafo de acordo com o modelo de gravata borboleta. Os resultados são apresentados na Tabela 10. Os dados mostram que os componentes SCC correspondem ao maior CFC de cada grafo. Esse componente contém cerca de 11% dos vértices do grafo `.br` e cerca de 27% dos vértices do grafo `.gov.br`. Este fato indica que o grafo `.gov.br` é melhor conectado.

O maior componente é o OUT, que em ambos os grafos inclui mais da metade dos vértices. Esse componente dirige os caminhamentos na Web para fora do SCC (ver Figura 1). Portanto, ao percorrer esse grafo, assim que este componente for alcançado, a navegação levará a um caminho sem volta ao SCC. Esse é, sem dúvida, um grande impedimento à conectividade da Web. Outro fato que chamou a atenção foi a quantidade de vértices no componente Tendrils no grafo `.br`, que também leva a caminhos sem retorno ao centro da rede, como no caso do componente OUT.

4.4. Cobertura e Limitações da Coleta

Uma questão importante para a análise da relevância dos resultados é quão representativa é a coleta realizada no contexto da Web brasileira. Em outras palavras, qual é a proporção de sítios coletados em relação ao universo de domínios registrados no Brasil. O número de domínios registrados muda diariamente, uma vez que domínios podem ser criados e cancelados continuamente. No início de abril de 2014, havia 3.386.165 domínios de primeiro nível (DPN) registrados no país, conforme mostra a Tabela 11. A mesma Tabela mostra o número de sítios contados na coleta da Web brasileira, em cada domínio. Em relação ao total oficial de sítios registrados no país, a coleta corresponde a 7,48%. Em relação ao DPN `.com.br`, que representa o maior percentual dos DPN (91,14%), a coleta alcançou 7,17% do universo. Finalmente, em relação à Web governamental, correspondente ao DPN `.gov.br`, a coleta alcançou 46,88%.

Tabela 11. Cobertura da primeira coleta em relação aos DPN registrados no país

DPN	Registro.br	Coleta	Percentual (%)
.com.br	3.086.226	221.263	7,17
.net.br	101.748	1.243	1,22
.org.br	48.220	13.048	27,06
.blog.br	7.639	528	6,91
.edu.br	2.473	850	34,37
.gov.br	1.361	638	46,88
outros	138.498	15.566	11,24
Total	3.386.165	253.136	7,48

A coleta de dados via *crawler* é bastante eficiente em termos da taxa de processamento mas impõe limitações intrínsecas ao seu objetivo, que é indexar sítios e páginas. Assim, o *crawler* realiza uma busca em profundidade sem exaurir todas as conexões de um sítio antes de passar a outro. Esse processo pode limitar o número de *links* encontrado. No entanto, não há outra maneira de percorrer a Web de forma eficiente. Uma vez encontrados os sítios, uma opção é coletá-los de forma completa para análise, tarefa que requer ainda mais recursos computacionais para ser cumprida.

5. Conclusão

Este artigo apresentou uma investigação sobre a conectividade da Web brasileira, com base em uma coleta ampla do domínio `.br`, e também da Web governamental brasileira (`.gov.br`). Os dados indicam que mais de 7% da Web brasileira foi coletada e analisada nesse trabalho. Os resultados indicam que a Web brasileira é uma rede composta por sítios muito conectados internamente. As conexões internas aos sítios ocorrem em proporção muito maior do que conexões entre sítios. Esse fato indica que os projetistas de sítios buscam estabelecer ligações entre conteúdos e que estão bastante cientes da importância de incluir *links* em suas páginas. No entanto, a ligação entre sítios é ainda tímida. O grau médio dos vértices é um pouco acima de dois, apenas uma unidade a mais do que o grau mínimo (um) para manter um vértice conectado ao grafo. Os vértices tendem a se conectar àqueles que possuem muitas conexões. O baixo grau médio encontrado nos grafos e os altos valores de seus diâmetros são evidências de baixa conectividade. Outro indício de baixa conectividade desta rede foi encontrado nas medições de sua estrutura topológica. No grafo `.br` os maiores componentes são os Tendrils e OUT, justamente os componentes que, ao serem alcançados em um caminhamento no grafo, resultam no fim da navegação, pois impedem o caminho de volta ao núcleo do grafo.

A partir dos resultados obtidos desta investigação podemos concluir que a Web brasileira é muito conectada localmente, porém sua conectividade global é baixa. Os resultados apresentados nos levam a crer na impossibilidade prática de percorrer toda a Web brasileira seguindo as ligações entre sítios. Sem os motores de busca não conseguiríamos alcançar boa parte da Web brasileira. Em última análise, as máquinas de busca funcionam como um vértice altamente conectado, a partir do qual os demais são alcançados. Uma sequência natural do trabalho é aumentar os recursos computacionais utilizados na coleta para obtermos coletas maiores. Outra sugestão é alterar as estratégias de coleta e avaliar os resultados. Outras métricas de conectividade em grafos podem ser também calculadas.

Agradecimento

Esta pesquisa foi apoiada pelo Instituto Nacional de Ciência e Tecnologia para a Web (INWEB - CNPq no. 573871/2008-6), com fomentos das agências CAPES, CNPq, Finep e Fapemig.

Referências

Barabási, A. L. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(509):1–11.

Barabási, A.-L., Albert, R., and Jeong, H. (2000a). Scale-free Characteristics of Random Networks: The Topology of the World-Wide Web. *Physica A*, 281:69–77.

Barabási, A.-L., Albert, R., Jeong, H., and Bianconi, G. (2000b). Power-law Distribution of the World Wide Web. *Science*, 287:1–2.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D. (2006). Complex networks: Structure and Dynamics. *Physics Reports*, 424:175–308.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph Structure in the Web. *Computer Networks*, 33:309–320.

Bush, V. and Wang, J. (1945). As We May Think. *Atlantic Monthly*, 176:101–108.

Chayes, J. (2013). Mathematics of Web Science: Structure, Dynamics and Incentives. *Philosophical Transactions of the Royal Society A*, 371:1–4.

Donato, D., Laura, L., Leonardi, S., and Millozzi, S. (2007). Graph Mining: Laws, Generators, and Algorithms. *ACM Computing Surveys*, 7(1):1–25.

Gomes, D. and Silva, M. J. (2003). A Characterization of the Portuguese Web. *3rd ECDL Workshop on Web Archives*, pages 1–14.

Google (2008). We knew the Web was big... `http://googleblog.blogspot.com.br/2008/07/we-knew-web-was-big.html`. Acesso em 29 janeiro 2014.

Jack, P. and Binns, A. (2012). Heritrix - Internet Archive Webteam Confluence. `https://webarchive.jira.com/wiki/display/Heritrix/Heritrix`. Acesso em 31 julho 2013.

Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tompkins, A., and Upfal, E. (2000). The Web as a Graph. *Nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 1–10.

Modesto, M., Álvaro Pereira Jr., Ziviani, N., Castillo, C., and Baeza-Yates, R. (2005). Um Novo Retrato da Web Brasileira. *Anais do XXVI SEMISH*, pages 2005–2017.

Pajek (2013). Pajek - Program for Large Network Analysis. `http://pajek.imfm.si/doku.php?id=pajek`. Acesso em 12 dezembro 2013.

Veloso, E. A., de Moura, E. S., Golgher, P. B., da Silva, A. S., Almeida, R. B., Laender, A. H. F., Ribeiro-Neto, B., and Ziviani, N. (2000). Um Retrato da Web Brasileira. *Anais do XXI SEMISH*, pages 1–10.

Vitali, S., Glattfelder, J. B., and Battiston, S. (2011). The Network of Global Corporate Control. `http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0025995`. Acesso em 19 junho 2013.

Wikipedia (2004). The World Wide Web. [Online; acesso em 2 de abril de 2014].