Ecossistema de Ligações da Blogosfera Portuguesa

José Luís Devezas

Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal

joseluisdevezas@gmail.com

22 de Março de 2010

Conteúdos

Introdução

O que é a blogosfera?

Definição

A blogosfera consiste no conjunto de todos os blogues e suas interligações.

A blogosfera é:

- Uma rede de blogues;
- Um conjunto de textos ricos em imagem e vídeo;
- Um conjunto de entradas cronologicamente ordenadas.

Estudar a blogosfera

O estudo da blogosfera pode focar-se:

- Na evolução da colecção;
- No conteúdo das entradas;
- Nos comentários;
- Na estrutura de ligações.

Enunciado de tese

Através da análise de ligações, identificar e caracterizar conjuntos de blogues, com o objectivo de provar que:

Afirmação

Existe um padrão consistente de variação de características dos blogues com a popularidade.

Conhecer a amostra

Colecção

- Disponibilizada pelo SAPO;
- Entradas escritas em português;
- Vários domínios, principalmente Blogues do SAPO e Blogger;
- Entradas entre 1 de Março de 2006 e 1 de Outubro de 2009.

Tecnologias

- Base de dados relacional `MySQL`.
- Base de dados chave $\Rightarrow$ valor `Berkeley DB`.
- Dialecto `GraphML` para representação de grafos.
- Linguagem e ambiente `R` para computação estatística e gráficos.
  - Biblioteca `ggplot2` para criação de gráficos.
  - Biblioteca `igraph` para manipulação de grafos.
- Linguagem `Perl`.
  - Extracção e selecção de dados.
  - Processamento e indexação de conteúdos.
  - Geração do documento `GraphML` e tabelas de entrada para o `R`.

Extracção e validação do conjunto de dados

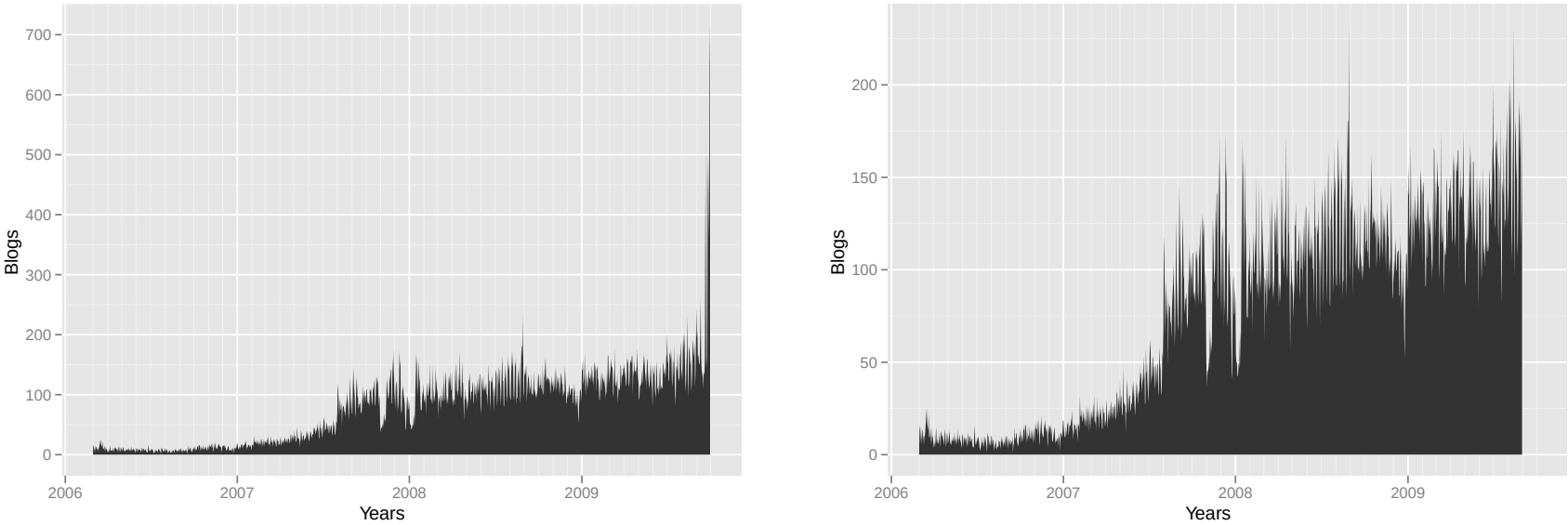
Dos 17 GB de registos são seleccionadas as entradas:

- Cujo domínio contém "blogs.sapo.pt";
- Datadas entre 1 de Março de 2006 e 30 de Setembro de 2009.

É feita a indexação de cada blogue no formato blogue ⇒ entradas:

```
blogue.blogs.sapo.pt =>
    http://blogue.blogs.sapo.pt/112.html|2008-02-01 23:45:32\t
    http://blogue.blogs.sapo.pt/342.html|2008-05-13 10:27:13\t
    http://blogue.blogs.sapo.pt/678.html|2008-11-11 11:13:27
```

Extracção e validação do conjunto de dados



Número de blogues criados por dia ao longo dos anos.

- Contagem do número de blogues e entradas criados por dia.
- Durante o mês de Setembro de 2009 observa-se um pico anormal.
- Uma verificação automática determina que 42% dos blogues desse mês não existem no mês seguinte.
- Setembro de 2009 é removido do estudo.
- Crescimento acentudado após a primeira metade de 2007.

Análise de ligações

Estrutura de dados

- Grafo dirigido para representar a rede de blogues.
  - Vértices ⇔ Blogues.
  - Arestas ⇔ Ligações entre os blogues.

    (provenientes de âncoras, imagens e conteúdo embebido no HTML das entradas)

- Vários atributos associados aos vértices e arestas.

	Atributo	Exemplo
Blogues	`name`	`blog.blogs.sapo.pt`
	`date`	`2007-10-11 16:22:57`
	`hostgraph.outdegree`	`50.077`
Entradas	`post.url`	`http://blog.blogs.sapo.pt/1046448.html`
	`post.date`	`2008-09-09 19:14:49`
	`post.wordcount`	`25`
	`post.charcount`	`216`
Ligações	`name`	`http://outro.blogs.sapo.pt/25856.html`
	`source`	`blog.blogs.sapo.pt`
	`target`	`outro.blogs.sapo.pt`

Informação captada no grafo de blogues.

Preparação dos dados

Passar da tabela de entradas, disponível na base de dados, ao grafo de blogues envolve:

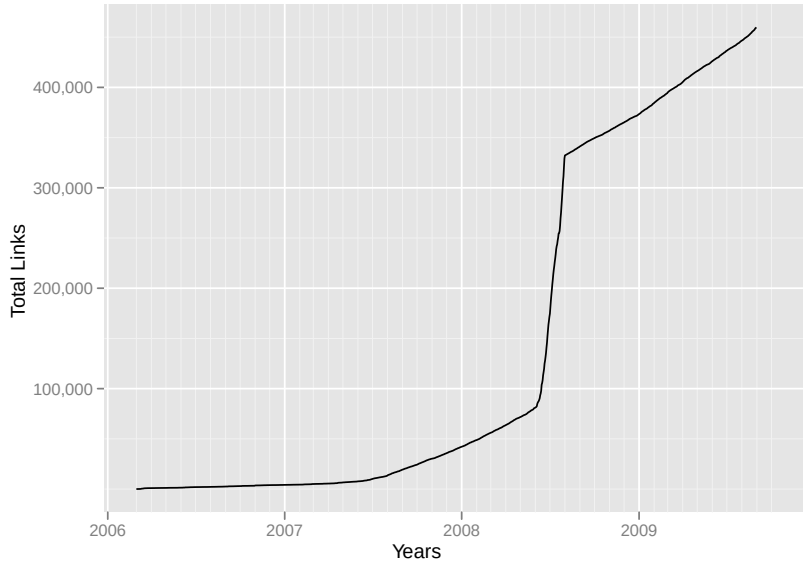
1. Extrair e indexar as ligações no formato ligação $\Rightarrow$ entradas

```
http://bit.ly/23a5b =>
    http://blogue.blogs.sapo.pt/112.html|2008-02-01 23:45:32|50|200\t
    http://outro.blogs.sapo.pt/1243.html|2008-05-13 10:27:13|19|101\t
    http://outro.blogs.sapo.pt/1122.html|2009-11-11 11:13:27|7|32
```

2. Agregar por domínio, contabilizando as ligações de entrada e de saída;
3. Remover domínios externos ao Blogues do SAPO e associar a data de criação a cada blogue;
4. Gerar um documento `GraphML` que represente a rede de blogues;
5. Carregar o documento `GraphML` no `R`, para ser analisado utilizando a biblioteca `igraph`.

Total de ligações, por dia, ao longo do tempo

- 459.737 ligações, extraídas de 72.591 blogues.
- Taxa média de crescimento mensal: 17,88%.
- Pico de utilização de ligações durante Junho e Julho de 2008.
- Resulta no aumento acentuado do número de ligações.



Agrupamento de blogues

- Blogues ordenados por popularidade.
- Número de citações como critério de classificação.
- Duas versões do grafo de blogues:

  Grafo original Uma ilustração crua da realidade da blogosfera $\Leftrightarrow$ quantidade;

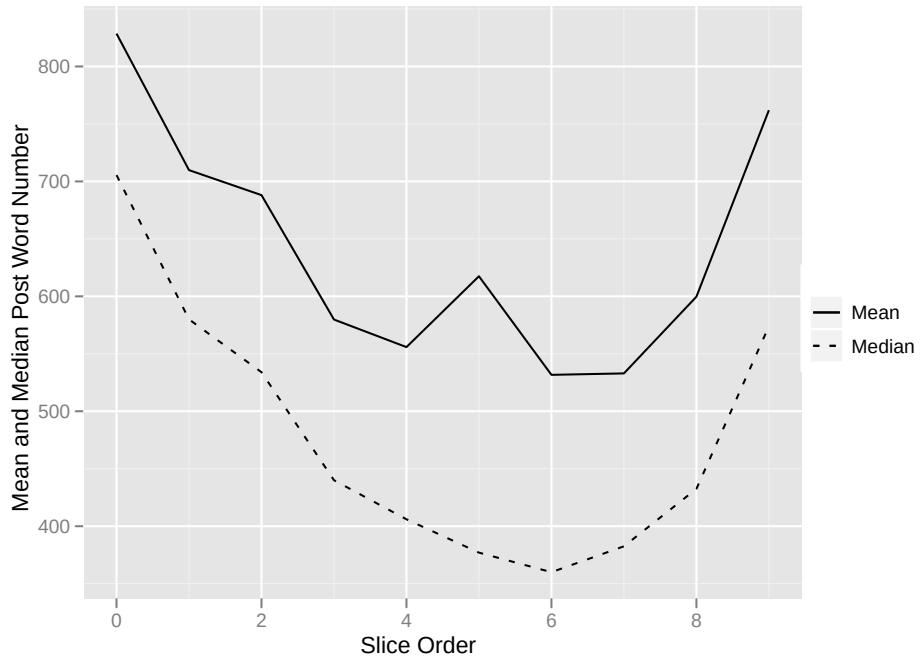
  Grafo simplificado Sem multiplicidade de arestas e auto-citações, ignorando nós com menos de duas ligações de entrada ou de saida $\Leftrightarrow$ variedade.

- Grafos de blogues partidos em fatias de 1.000 blogues.
- Análise da evolução do valor médio e mediano de várias características, para fatias progressivamente menos populares.

Número de palavras por entrada, para o grafo simplificado

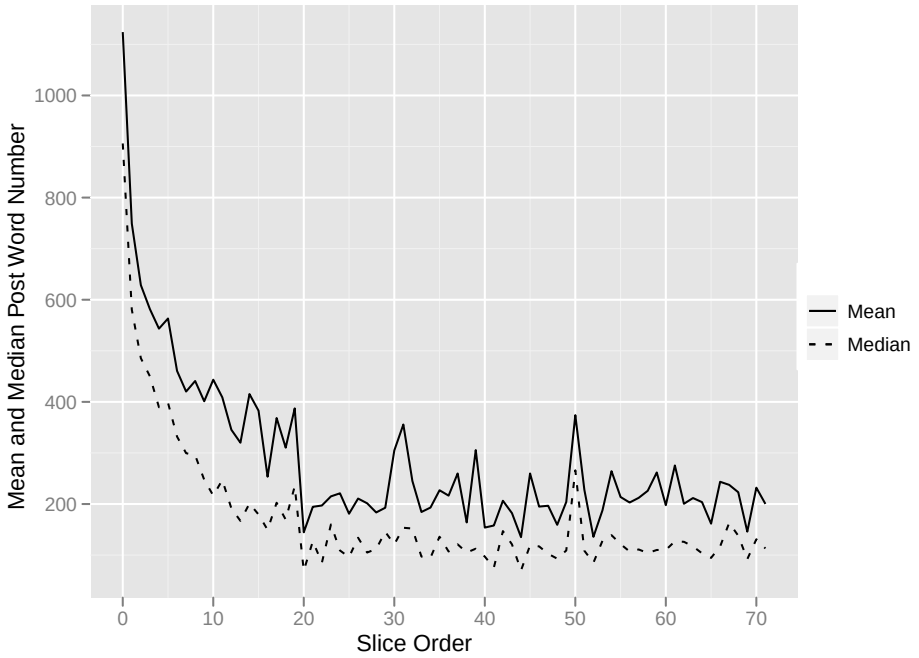
- Eixo dos XX: ordem da fatia — a fatia 0 representa o conjunto dos 1.000 blogues mais citados e a fatia 9 representa os 1.000 blogues menos citados.
- Eixo dos YY: média ou mediana do número de palavras por entrada nos blogues da fatia.

Ordem	Média	Mediana
0	829	706
6	532	360
9	762	574



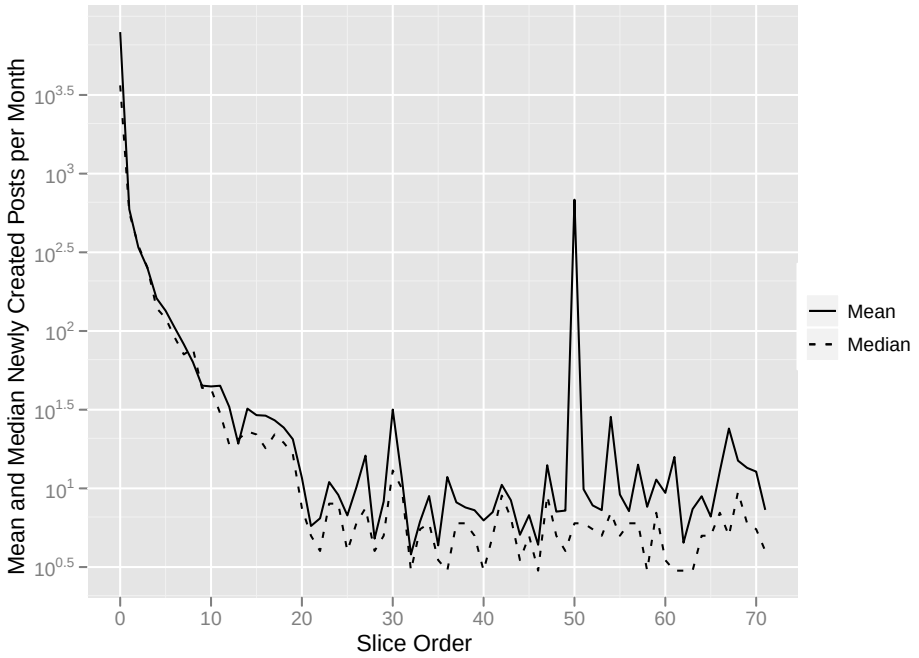
Número de palavras por entrada, para o grafo original

- Blogues mais citados: média de 1.124 palavras por entrada.
- Blogues restantes: média entre 135 e 749 palavras.
- Decréscimo constante, mas não muito acentudado.



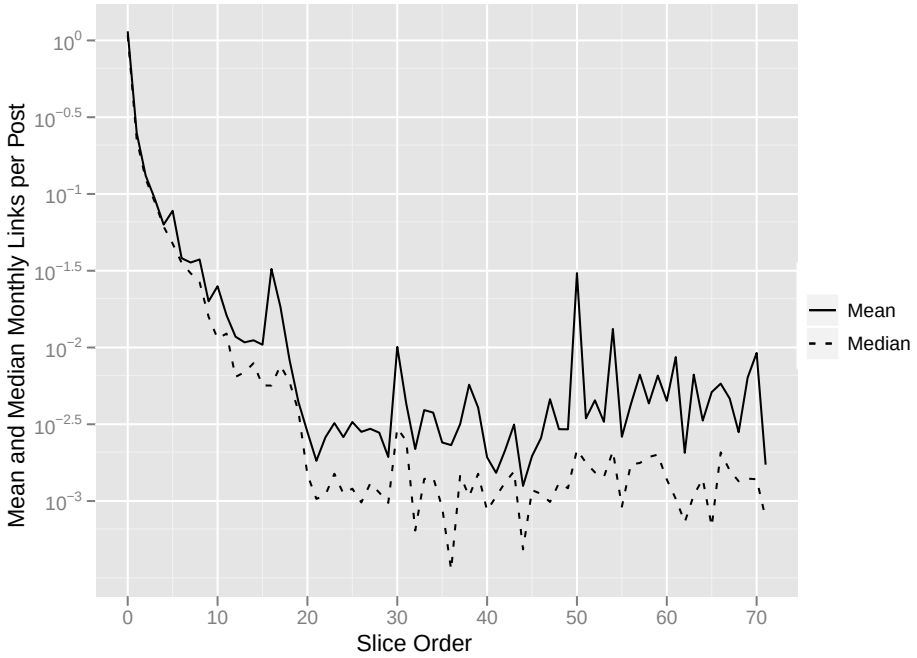
Número de entradas criadas por mês, para o grafo original

- Blogues mais citados: média de 7.934 novas entradas por mês.
- Outras blogues mais citados: média superior a 100 — 594 para a segunda fatia mais citada.
- Blogues menos citados: média entre 5 e 30 novas entradas mensais.



Número mensal de ligações p/entrada, para o grafo original

- Blogues mais citados são os que ligam mais a outros blogues.
- No entanto, em geral, as entradas não contêm um grande número de ligações.
- A fatia mais citada tem em média 1,15 ligações por entrada.
- As restantes fatias têm uma média inferior a 0,25 ligações por entrada.
- Na fatia menos citada é utiliza apenas 1 ligação por cada 1.000 entradas!



Conclusões

Principais contribuições

- Agrupamos blogues utilizando como critério de popularidade o número de citações.
- Blogues populares têm um comportamento distinto dos blogues menos populares.
- Conforme passamos dos blogues mais populares para os menos populares, observámos um padrão de decréscimo:
  - Na frequência de criação de entradas;
  - No número de ligações de saída;
  - No tamanho das entradas.
- Conclui-se que existem efectivamente grupos de blogues com características distintas.

Trabalho futuro

- Estudar a evolução da popularidade dos blogues.
  - O que influencia um blogue a tornar-se popular.
  - Como evolui a classificação dos blogues mais populares.
  - Como evoluem as suas características.
- Estudar as comunidades portuguesas de blogues.
  - Analisar algoritmos de detecção de comunidades.
  - Identificar o factor de união dos elementos das comunidades.
  - Identificar os blogues centrais de cada comunidade.

Fim

Questões?

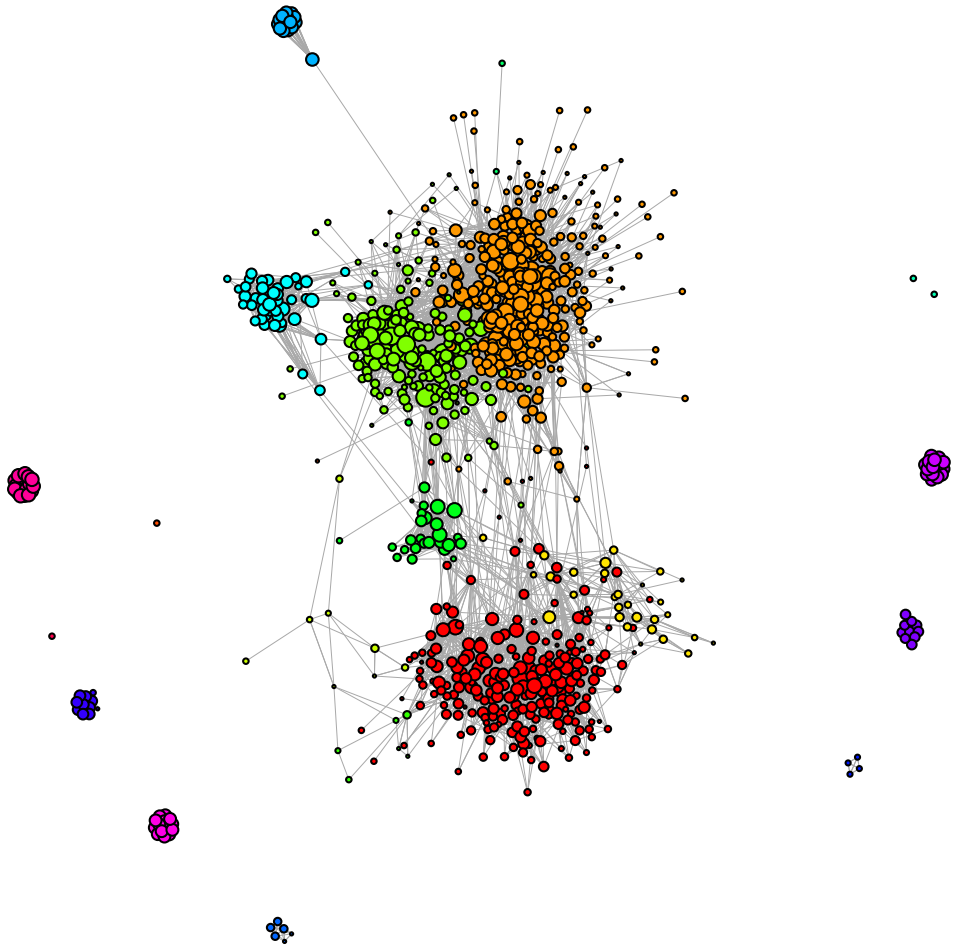
Apêndice

Apêndice A.1

Metáfora de ecossistema

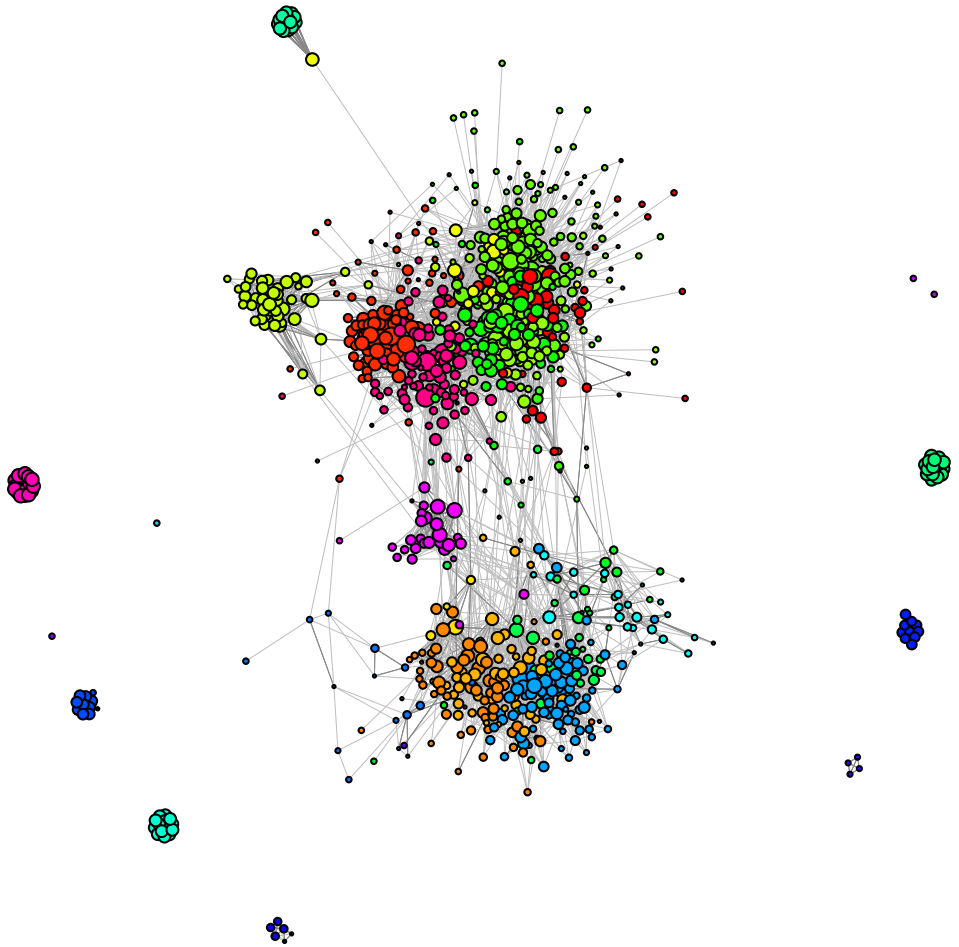
A blogosfera pode ser vista como um ecossistema em que os blogues são considerados organismos que interagem entre si, interligando-se por meio de hiperligações, no ambiente da *World Wide Web*.

Apêndice B.1



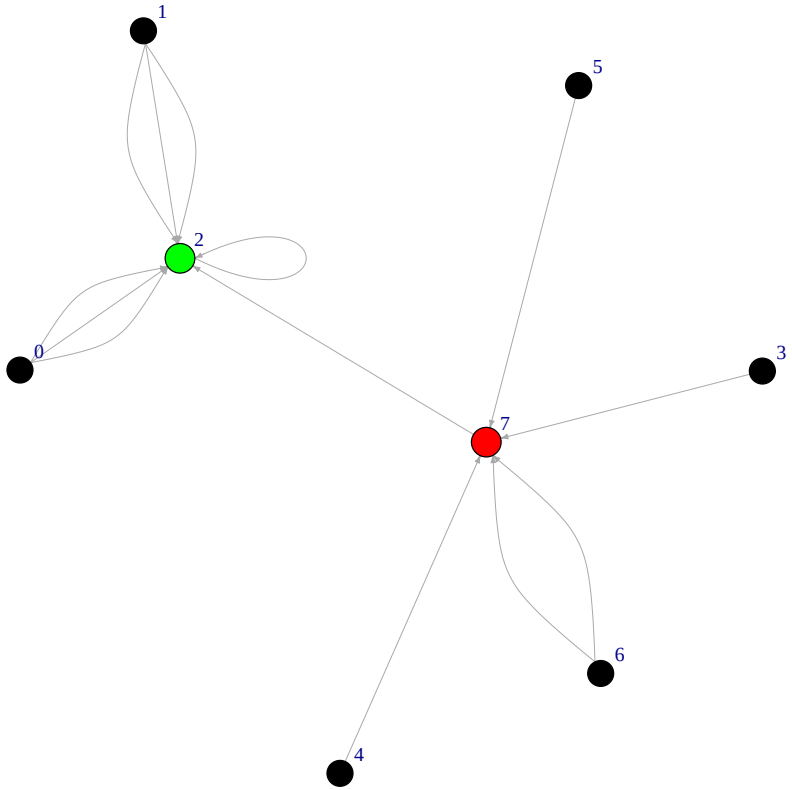
Detecção de comunidades utilizando o algoritmo *Walktrap*.

Apêndice B.2



Detecção de comunidades utilizando o algoritmo *Leading Eigenvector*.

Apêndice C.1

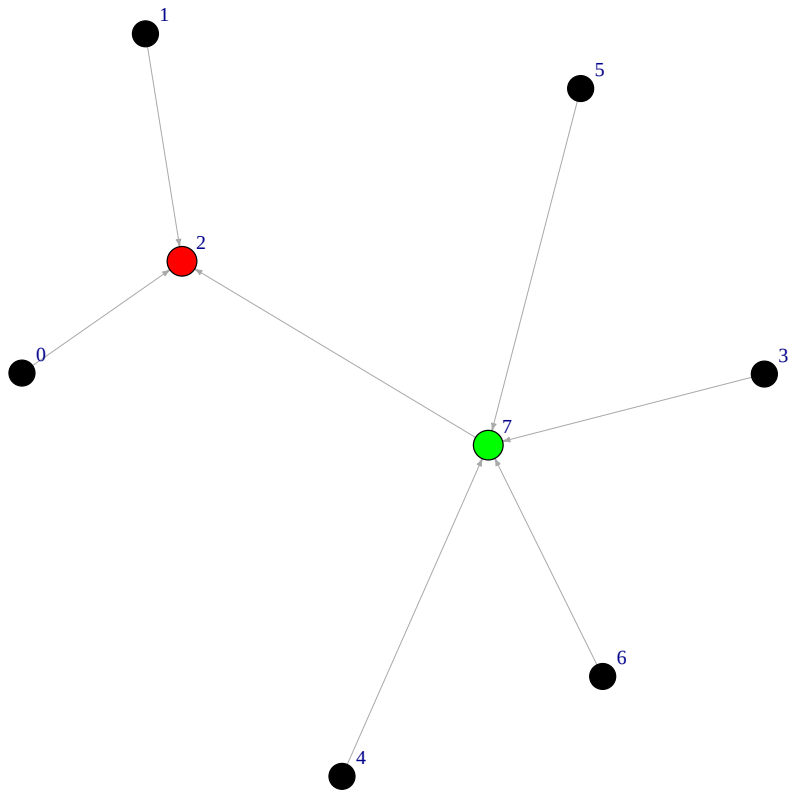


Amostra do grafo de blogues antes da simplificação.

Classificação

Nesta amostra, o nó 2 é citado 8 vezes e o nó 7 é citado 5 vezes. O nó 2 é considerado o mais popular devido à quantidade de ligações que apontam para ele.

Apêndice C.2

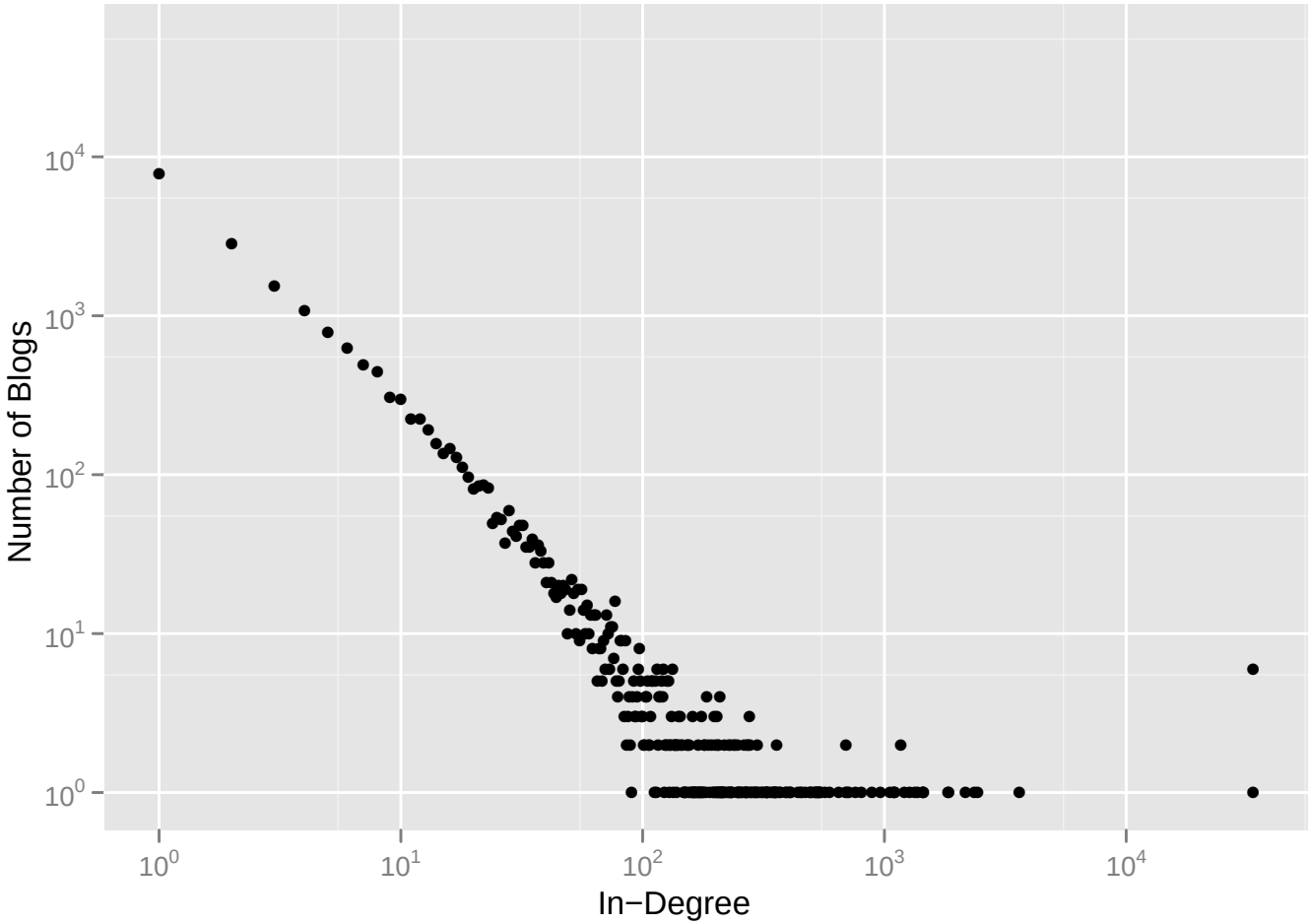


Amostra do grafo de blogues após a simplificação.

Classificação

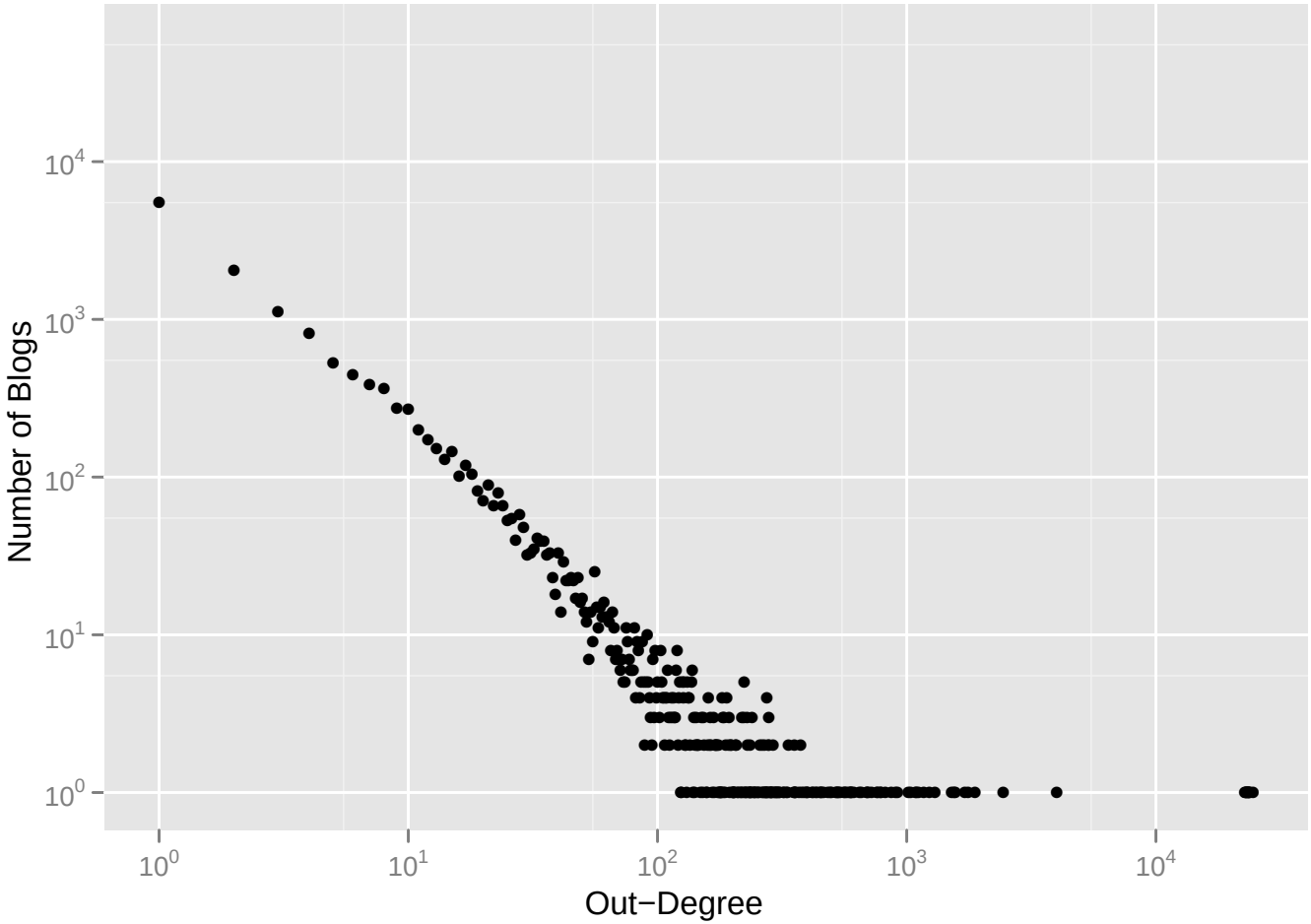
Nesta amostra, o nó 2 é citado 3 vezes e o nó 7 é citado 4 vezes. O nó 7 é considerado o mais popular devido à variedade de ligações que apontam para ele.

Apêndice D.1



Distribuição do grau de entrada do grafo de blogues.

Apêndice D.2



Distribuição do grau de saída do grafo de blogues.

Apêndice E.1

	Grafo original	Grafo simplificado
Vértices	72.591	10.937
Arestas	459.737	48.399
Densidade	8,72e-5	4,05e-4
Reciprocidade	0,31	0,15
Diâmetro	20	19
Ligações por blogue (média)	6,33	4,43
Ligações por blogue (mediana)	0	2

Apêndice F.1

- Estudar o comportamento dos blogues dependendo da popularidade permite:
  - Desenvolver estratégias de mercado para publicitar blogues profissionais;
  - Melhorar o serviço de blogues, adaptando-o aos utilizadores mais activos.
- Estudar as comunidades de blogues permite:
  - Localizar um blogue na sua comunidade;
  - Fornecer ao autor do blogue a sua vizinhança;
  - Determinar os blogues mais centrais de cada comunidade.

Apêndice G.1

- The Most Important Blogging Analysis Ever:
  - http://www.viperchill.com/important-blogging-analysis/
- The Secret to Building a Popular Blog (and Getting Tons of Readers):
  - http://www.doshdosh.com/the-secret-to-building-a-popular-blog/
- Why is Link Popularity Important?
  - http://blogs.siliconindia.com/uniquesofts2010/Why_is_Link_Popularity_Important-bid-gJ1zb05l46050252.html