Studying Blog Features over Link Popularity

José Luís Devezas[‡], Cristina Ribeiro[†‡], Sérgio Nunes[‡]

INESC-Porto[†]
DEI, Faculdade de Engenharia, Universidade do Porto[‡]

FEUP
Universidade do Porto
Faculdade de Engenharia

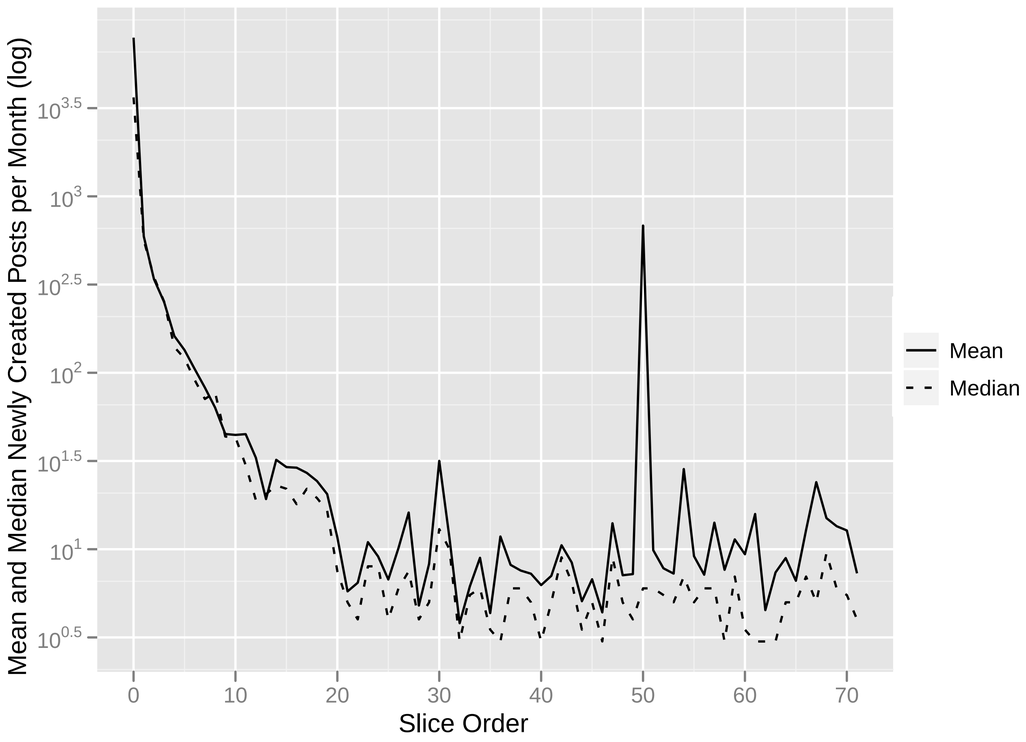

Figure 1: Newly created posts per month.

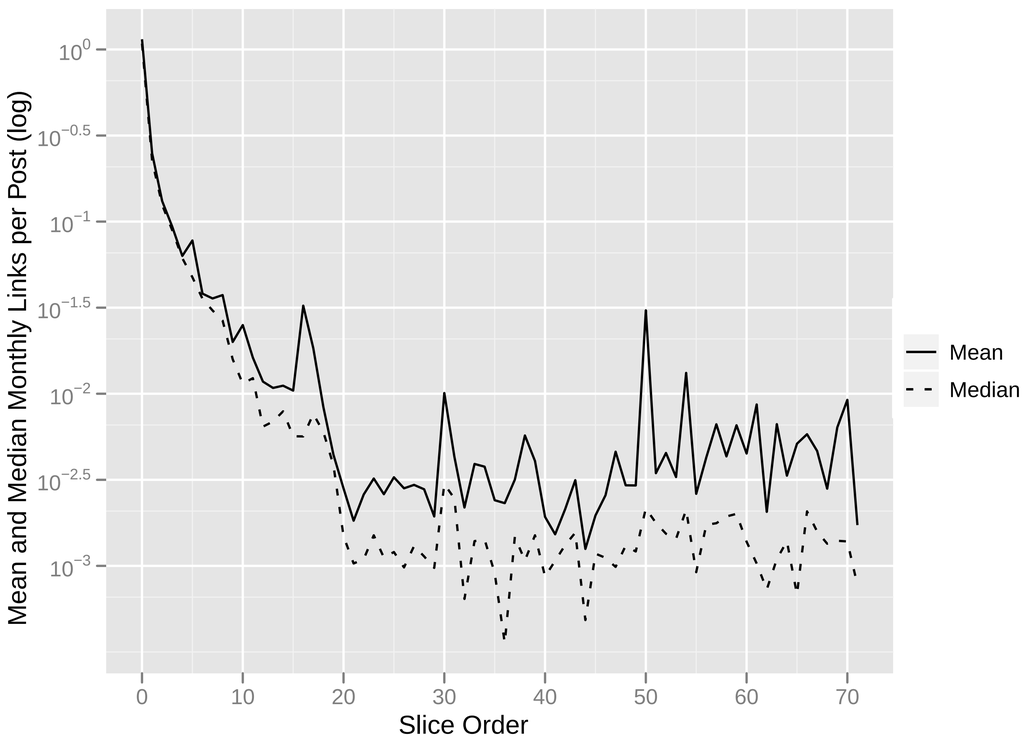

Figure 2: Monthly number of out-links per post.

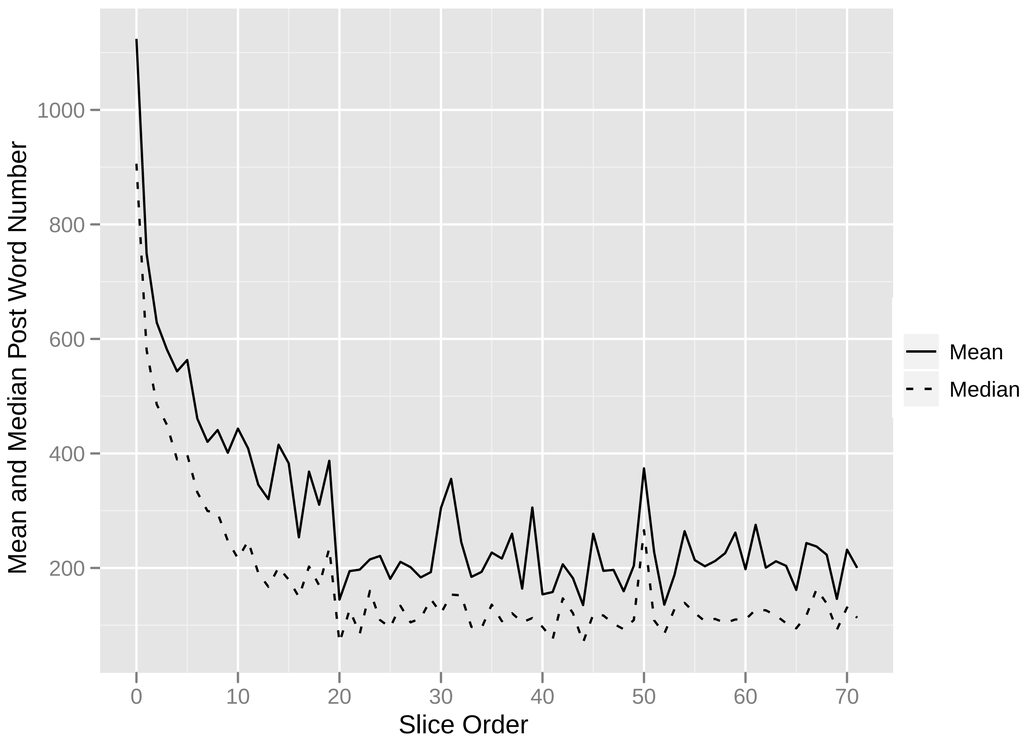

Figure 3: Number of words per post.

Collection

- Provided by SAPO, a portuguese ISP and blogging service.
- **100,000+ blogs**, with over **2 million posts**, written in **portuguese**.
- Several blog domains, mainly **SAPO Blogs** and **Blogger**.
- Dated from March 1st 2006 to October 1st 2009.

- Data set built from a 17 GB table, by selecting posts:
  - Whose domain contained ".blogs.sapo.pt".
  - Dated between **March 1st 2006** and **September 30th 2009**.
- We study more than **70,000 blogs**, with over **400,000 links**.

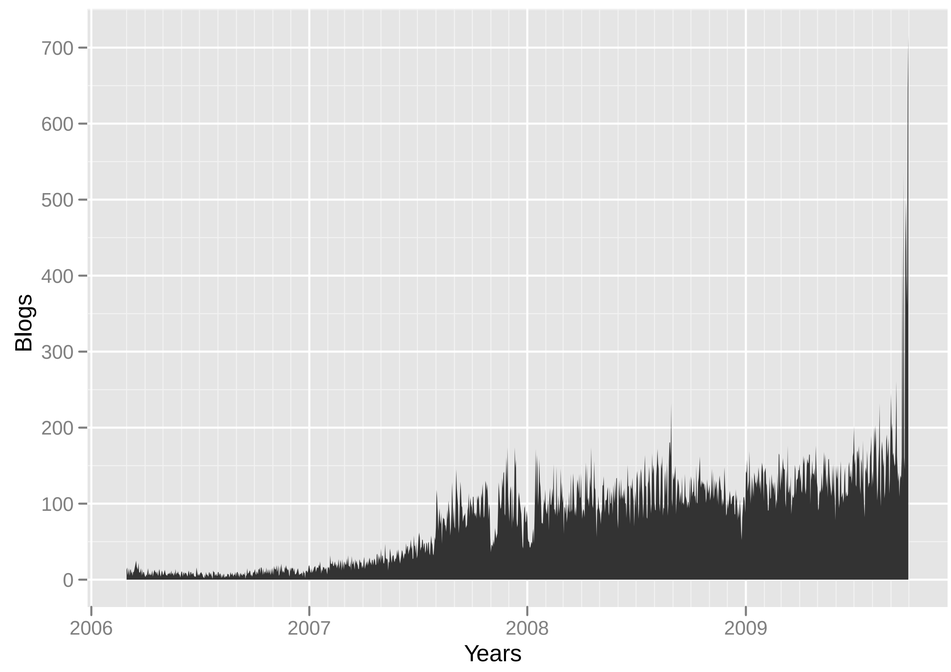

Figure 4: Newly created blogs per day.

Data Validation

- Abnormal peak in September 2009 (Figure 4).
- **Automatic verification:** 42% of the blogs in September 2009 do not exist in October 2009.
- **Conjecture:** Splog cleaning processed has been applied to the collection, but not yet to September 2009.
  Decision: Remove September 2009 from the study.

Data Preparation

1. Extract URLs from HTML anchors, images and embedded resources, storing them in a Berkeley DB, as URL $\Rightarrow$ {posts}.
2. Aggregate by hostname, removing the domains external to SAPO Blogs.
3. Generate a GraphML document and load the graph into R to be studied.

Results

- Studying the evolution of several features, for slices of 1,000 blogs, ordered by number of citations, reveals a decreasing pattern in:
  - Post creation frequency (Figure 1).
  - Number of out-links (Figure 2).
  - Number of words of the posts (Figure 3).
- Popular blogs have distinct behaviors when compared to less popular blogs.

FEUP Universidade do Porto Faculdade de Engenharia
sapo.pt
FCT Fundação para a Ciência e a Tecnologia MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR
INESCPORTO INSTITUTO DE ENGENHARIA DE SISTEMAS E COMPUTADORES DO PORTO