On Bias-Free Crawling and Representative Web Corpora

(supported by the German Research Council (DFG), grant SCHA1916/1-1)

Roland Schäfer
German Grammar, Freie Universität Berlin
`http://rolandschaefer.net/`

Web Corpora and Crawling

- Web corpora are (virtually) **always based on crawls**.
- COW, LCC, UMBC WebBase, WaCky,. . .

Biber (1993): ***"[T]heoretical research should be prior in corpus design"*... but is it really with ordinary crawled web corpora (and Google-scraped corpora)?**

Breadth-First Bias (Achlioptas et al. 2005; Kurant et al. 2010; Maiya and Berger-Wolf 2011)

- Breadth-first search (BFS): biased towards **in-degree**
- Bias **cannot be corrected** post-crawl!
- This is a problem – unless we believe that:
  1. High in-degree means high relevance. **Nonsense!**
  2. Google knows what's good for us.
     Is this what Biber (1993) had in mind?

Solutions (based on Henzinger et al. 2000; Rusmevichientong et al. 2001)

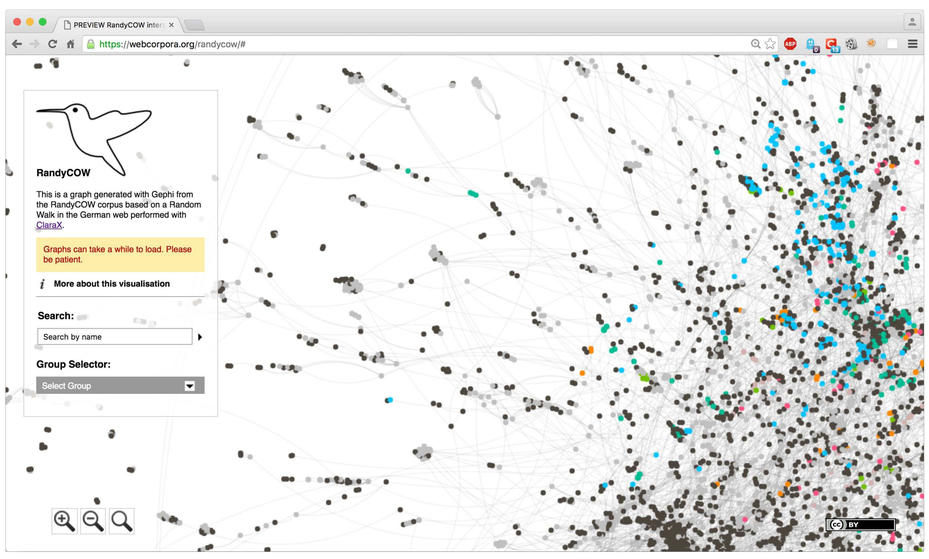
- Use (slow) **Random Walks** (RW) and . . .
- . . . **correct PageRank bias** post-crawl.
- Only for small reference corpora

Representative corpora in a purely sampling-theoretical/statistical sense!

One Goal: Linguistic Web Characterization

- Assessment of the true composition of the web
- Linguistic characterization of **"web of web hosts"** by **lexico-grammatical** features and **topics**
- COReX feature set (with IDS Mannheim)
- COReCO topic domain classification (with IDS)
- **Basis for stratification of larger web corpora**

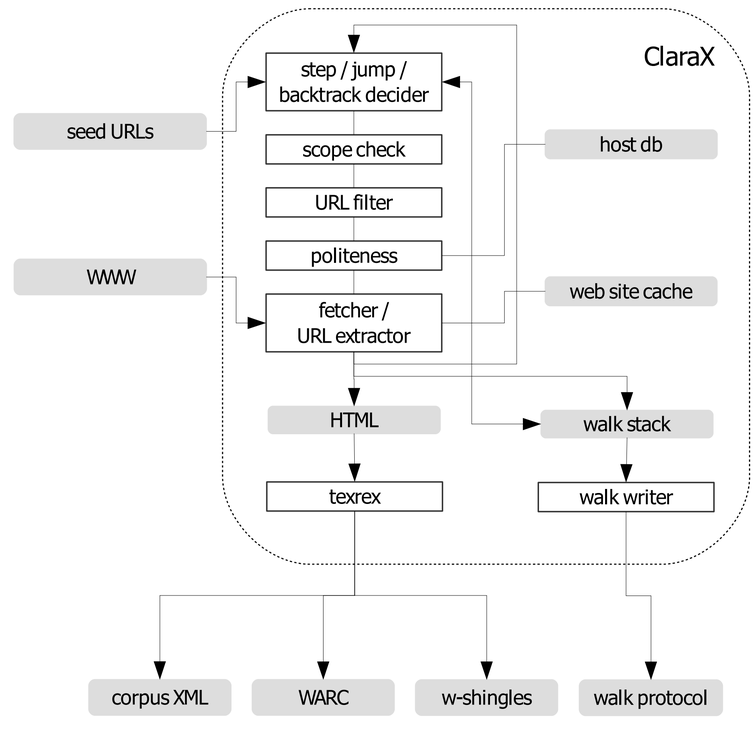
RandyCOW: Users can explore the (reasonably bias-free) web corpus graph by coloring nodes depending on distributions of linguistic features.



This is a 20% functional preview. Major TODO: Better/selectable graph layouts. Based on **sigmajs.org**, prepared with **Gephi**.

ClaraX: A Random Walker

- Fully-featured **Random Walker** (i. e., not a "crawler")
- ***texrex*** post-processing integrated (Schäfer et al. 2012 etc.), derived from **HeidiX**
- **2-clause BSD license**
- `https://github.com/rsling/texrex`



First Two Experiments

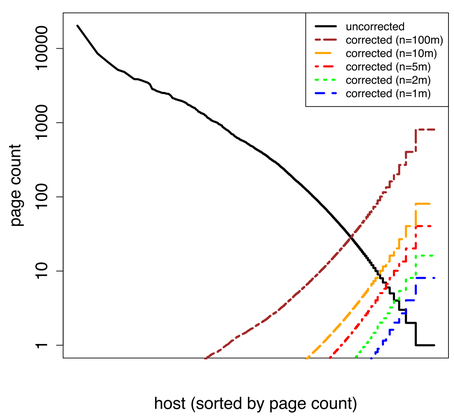
Baseline experiments in German-speaking web:

1. Follow **any link** (true page-wise RW)
2. Jump **from host to host** (host-wise RW) with Henzinger-style bias correction post-crawl

Exper.	Runtime	Steps	Hosts	St./Host
1	12.75d	1,093,047	1,227	890.83
2	25.36d	2,090,443	204,053	10.25

Steps	Host
91,442	www.vsw-news.de
40,806	pauls-blog.over-blog.de
35,787	fielders-choice.de
34,411	www.my-bikeshop.de
34,091	www.bremer-treff.de
24,769	www.deutscher-werkbund.de
24,114	www.vau-niedersachsen.de
24,096	www.icony.de
22,299	www.discover.de
20,093	www.dewezet.de

The 10 longest RW segments spent on a single host during the first experiment



Number of pages (y) visited in the second experiment per host (x), sorted in decreasing order, and the theoretically expected document counts when applying Henzinger's rejection sampling method depending on the targeted bias-reduced corpus size, given as n; log-log axes

- We need **longer walks** (6-month walk running).
- We have to experiment with **less aggressive bias correction** (incl. graph simulation).