1. Introduction

Crawling, scraping, visualising Australian political blogs:
- Based on growing list of blog feeds
- Custom-built content scraper
- Distinction between post content and other material

Timeframe:
- November 2007 to January 2008 — ~250 feeds
  (for Australian federal election on 24 November 2007)
- Second stage ongoing since January 2008

2. Method

- Identify blogs, subscribe to RSS feeds
- Follow RSS feeds, identify new posts
- Scrape entire blog page
- Isolate salient content — blog post only (plus comments?)
- Extract post text and links
- Map topics and link networks

Advantages over previous approaches:
- Better data quality:
  - excluding blog headers / footers / sidebars / ads / blogrolls
- More than links:
  - capturing content for automatic and manual analysis
- Not retrospective:
  - tracking and capturing activity as it happens
- Diachronic, diatopic:
  - comparing activity across timeframes and blogs / blog clusters

3. Results

- clear distinctions in topical focus of specific blogs
- moderate clustering tendencies
- links from mainstream online media very rare
- mainstream media blogger-columnists in a separate category
- left-leaning blogs more active than right-leaning blogs
- weekly and seasonal activity variations
- activity triggered in part by external events

4. Conclusions

- successful proof of concept for our tools
- further tweaking of data outputs required
- able to provide weekly/monthly pointer to ‘hot’ topics/sites
- need to operate for 6-12 months to establish baseline data
- can be expanded to tracking activity in other user-led content creation spaces (Twitter, Facebook, YouTube, ...)

5. References


More than links:
Automated and manual analysis of textual content — e.g. using Leximancer:

Diachronic comparisons:
Analysis of activity per timeframe — e.g. weeks before and after 2007 election:

Data quality:
Only relevant, salient content is selected by the scraper.

social policy

economy

parties and leaders