Nested Palindromes in Clickstream Data

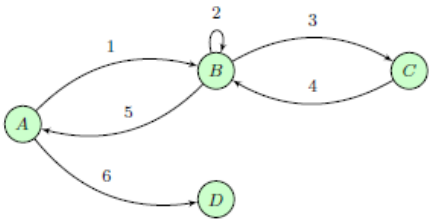
By: Speiser, Antonini, Labbi, and Sutanto
Presented By: William Garrard

Clickstream

- Pages you go through when visiting a site
  - Sequence of URL's for single device, user, session
- Inactivity of > 30 minutes denotes a new visit

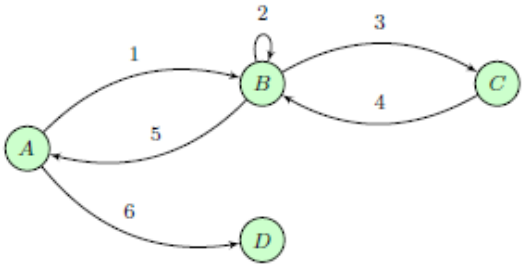
Palindrome

- Sequence with central symmetry
  - In words: racecar, kayak
  - In page names: ABBCBAD
- Refreshes count as a reflexive arc, backtracks count as a pair of opposite arcs



Nested Palindrome

- Three repetitions
  - 2xB, 1xA
- BB, BCB are palindromes



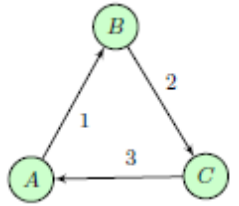
Compressed Palindrome

- Replace first-ending palindrome with outer symbol
  - Continue until out of palindromes
- If no repetitions remain, they are all *accounted for*

sequence	1st palindrome "compressed"
$ABBCBAD$	$ABCBAD$
$ABCBAD$	$ABAD$
$ABAD$	AD
AD	AD

Compressed Palindrome

- Unaccounted sequences exist
  - Repetitions , but no palindromes
- ABCA
- These are a minority
- Must be length >=4



Dataset

- Obtained via 'sessionization' of server logs
- 3 sets, named Blue, Orange, and msnbc.com
  - IBM.com and possibly a competitor?
- Did not count visits over 20 pages long

statistic	Blue	Orange	MSNBC
sample size	$3.1 \cdot 10^5$	$3.2 \cdot 10^5$	$9.9 \cdot 10^5$
distinct items	1822	8822	17
visits over length 20 (removed)	0.16%	0.53%	2.8%

Evidence of Nested Palindromes

- Unaccounted sequences are rare

unaccounted visits among all visits	0.86%	1.6%	5.4%
unaccounted visits among visits of length ≥ 4 containing repetitions	14%	18%	8.3%

- Palindromes are very common

proportion of palindromic repetitions	95%	92%	96%

Applications: Pre-Processing

- Addition of flags within sequences aids in analysis
  - ABBCBAD → AB**X**BC**Y**B**Z**AD
- ABCYB is frequent
  - Equivalently, ABCB would be frequent
  - Prevents counting ABCDEB, other combinations

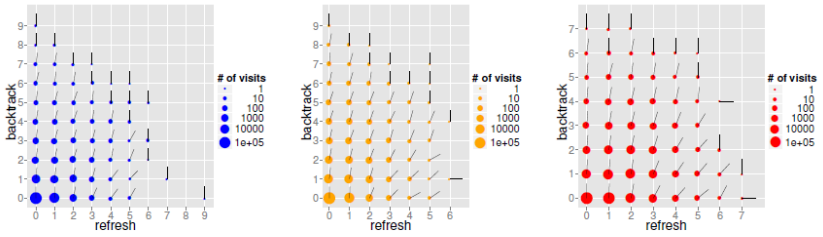
Applications: Pre-Processing

- Finding frequent patterns, how many contain added symbols?

frequent closed partial orders containing added symbols	193/418 (46%)	313/339 (92%)	317/437 (73%)

- Can now identify the exact pages for which refreshes and backtracks are occurring

Refreshes and Backtracks



(a) Blue dataset (b) Orange dataset (c) MSNBC dataset

Blob size: number of visits
Spokes: proportion of accounted visits
-Vertical = 100%, Horizontal = 0%

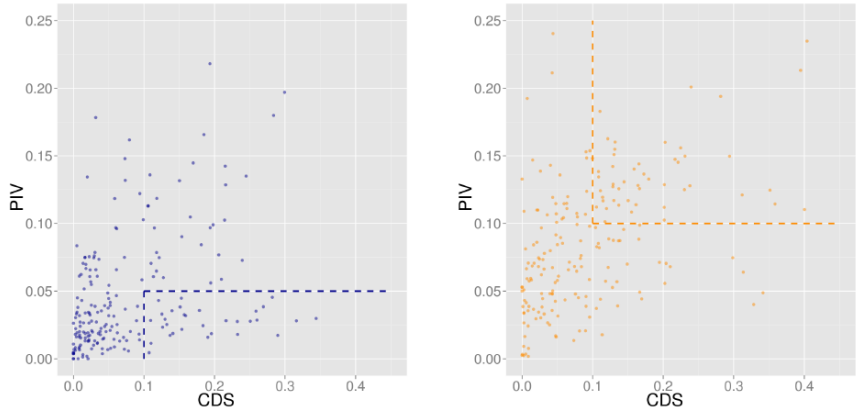
Applications: User Experience

- Calculate *support* of palindrome patterns in a visit
  - If given pattern occurs n>=1 times, support = 1
- $sup_{re}(x)$ = visits with a refresh of x
- $sup_{bf}(x)$ = visits with a backtrack from x
- $sup_{bt}(x)$ = visits with a backtrack to x
- Refresh Rate: $REF(x) = sup_{re}(x) / sup(x)$
- Cul-de-sac Rate: $CDS(x) = sup_{bf}(x) / sup(x)$
- Pivot Rate: $PIV(x) = sup_{bt}(x) / sup(x)$

Applications: User Experience

- All ratios range between 0 and 1
- Can be good or bad depending on type of page that x is
- High refresh rate?
  - Good for breaking news/dynamic page
  - Bad for static page, should trigger examination
- High CDS rate?
  - Good for help page
  - Bad for portal/hub/index page
- High Pivot Rate?
  - Good for portal/hub/index page
  - Bad for workflow pages
- High CDS and Pivot?
  - Indicates page corridor

Blue and Orange Results



- Blue: many dead ends
- Orange: many corridors

Conclusions

- Real clickstream data experiences palindromicity very often
- Easily analyzable with simple algorithms
- Direct real-world impact from analysis

Limitations

- Limited to first-order palindromes
  - ABCBA does not produce (C,B,A) backtrack
- Should have found better data than msnbc.com

My Take

- Lack of corroboration with users for validate results
  - Each situation is domain specific
- Overall a useful high level analysis of websites
- Utilized on my own data

Questions?