Archiving the Relaxed Consistency Web

Zhiwu Xie, Herbert Van de Sompel, Jinyang Liu, Johann van Reenen, and Ramiro Jordan

October 31, 2013
CIKM 2013, San Francisco
Web Archiving

- Preserve the ephemeral web for long term use
- Document the common web experience, or the “collected memory” as cultural heritage
- Internet Archive, national libraries and archives, other non profit organizations
- Not only for history buff, but frequent weapon in legal battles
- Mainly crawler based
- Very effective on static, document web, less effective on highly dynamic web, faces challenges from the relaxed consistency web
Relaxed Consistency Myth

- Inconsistency is barely noticeable, therefore harmless.
The Trial of Bradley Manning

- On Tuesday June 18, 2013, the military prosecutor of the Pfc Manning’s trial wanted to admit two pieces of evidence.

- Both are retrieved directly from Twitter and then compared against their Google Cache copies to show they are authentic.
Prosecutor Alleged:

• “…shows Manning conspired with WikiLeaks…”
• “When there’s evidence of a plan, that evidence can also be used as proof of subsequent act,” Von Elten claimed. It’s “evidence they will compromise classified information in the future going forward.”
Defense Lawyer Argued:

- The “plan or state of mind of WikiLeaks has nothing to do with Pfc. Manning. They can plan or do whatever they want. That doesn’t affect Pfc. Manning.”

- “Did he actually see any of this? Did he ever actually see it?” asked Tooman. “There has to be a connection with Pfc. Manning and these documents and there’s not a connection between Pfc. Manning and these documents.”
Court Admitted the Tweets

• … these tweets were both relevant [to the charges] …
• The government has provided no evidence that Manning saw either of these tweets, but Judge Lind ruled that they were relevant as circumstantial evidence, due to their timing and public availability, and the fact that Manning was known to have searched Intelink (the military’s Google) for ‘WikiLeaks.’
Assumptions:

- Manning had a Twitter account and was following WikiLeaks
- The prosecutor was able to obtain an access log showing Manning had retrieved his Twitter timeline a few seconds after the timestamps of the WikiLeaks tweets
Question:

• “Did he actually see any of this? Did he ever actually see it?”
Not Your Grandfather’s Web Anymore

- Large number of users
- Personalized to the extreme
- Near real-time event updates
What is Consistency?

- A consistent system, even built on distributed machines, guarantees an illusion of a total order in which concurrent events can be observed and interpreted as happening on a single machine
- Guarantees common experience
Relaxed Consistency

- A relaxed consistent system is allowed to have a period of “inconsistency window” during which a global order cannot be established
Why Relaxed Consistency?

- CAP Theorem
- Better user experience: latency and scalability over consistency
How to Relax Consistency?

- Relaxed consistency data store
- NoSQL data model: no transaction guarantee
- Inconsistent caching
- ...
- Compound of all the above
Inconsistency Window

- Within seconds at the data store
- But at the full web application level?
A Controlled Experiment

• A simplified Twitter-like feed following application
• A realistic, mid-size following network
• A benchmarking workload representing the normal working condition of a social network
• Detect inconsistency
Observable Inconsistency

- User responses that conflict with each other
- Ignore inconsistency that have not been consumed by the end users
Inconsistency Example

WikiLeaks

Manning

Jone Doe

WikiLeaks posts a tweet

John Doe saw it

Private Manning did not see it at a later time

Time

$t_i$

$t_{i+1}$
## Workload

<table>
<thead>
<tr>
<th></th>
<th>PNUTS</th>
<th>This</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of producers</td>
<td>67,921</td>
<td>67,882</td>
</tr>
<tr>
<td>Number of consumers</td>
<td>200,000</td>
<td>196,283</td>
</tr>
<tr>
<td>Consumers per producer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>15.0</td>
<td>13.38</td>
</tr>
<tr>
<td>Zipf parameter</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>Producers per consumer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>5.1</td>
<td>4.63</td>
</tr>
<tr>
<td>Zipf parameter</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>Per-producer rate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>1/hour</td>
<td>1/hour</td>
</tr>
<tr>
<td>Zipf parameter</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>Per-consumer rate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>5.8/hour</td>
<td>5.8/hour</td>
</tr>
<tr>
<td>Zipf parameter</td>
<td>0.62</td>
<td>0.62</td>
</tr>
</tbody>
</table>
Experiment configuration

Load Generator
M2.2xlarge
Httpperf + autobench

Frontend Server
m1.large
Django + httpd + PostgreSQL

DynamoDB
1000 Units of Read/Write Capacity

POST Only

GET Only

GET Only
Results: Inconsistency Rate

- 6.27% contain observable conflicts
Results: Inconsistency Window

- average inconsistency window is 823 seconds
- The promise of eventually consistent holds
Implications to Web Archiving

- No guarantee for archiving the common experience
- Existing archives may have been polluted with inconsistency
- Archiving crawler’s behavior may not affect the inconsistency level it observes
- Popular tweets are more prone to inconsistency
Key Takeaways

• Inconsistency can be more severe than expected.
• Even with strengthened intelligence, the government may still not have a watertight case against Manning.