Challenges in Archiving the Relaxed Consistency Web

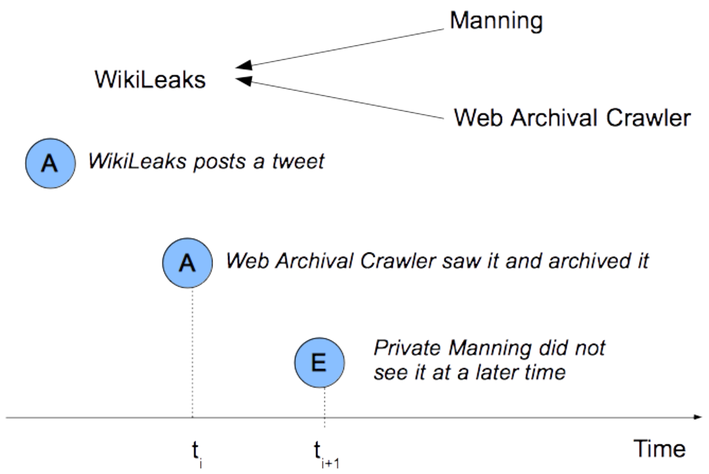
[1,4]Zhiwu Xie, [2]Herbert Van de Sompel, [3]Jinyang Liu, [4]Johann van Reenen, and [4]Ramiro Jordan
[1]Virginia Tech, [2]Los Alamos National Lab, [3]HHMI Janelia Research Campus, and [4]University of New Mexico

Consistency? What 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. Consistency guarantees common experience. A relaxed consistent system, on the other hand, is allowed to have a period of "inconsistency window" during which a global order cannot be established.

Today, the relaxed consistency technologies are becoming prevalent in many if not all leading web portals, news aggregators, and social networks. It is often hard to pinpoint which website uses what technology to relax consistency to what level unless disclosed by their technical team.

Inconsistency Degrades Archival Quality



New Phenomenon

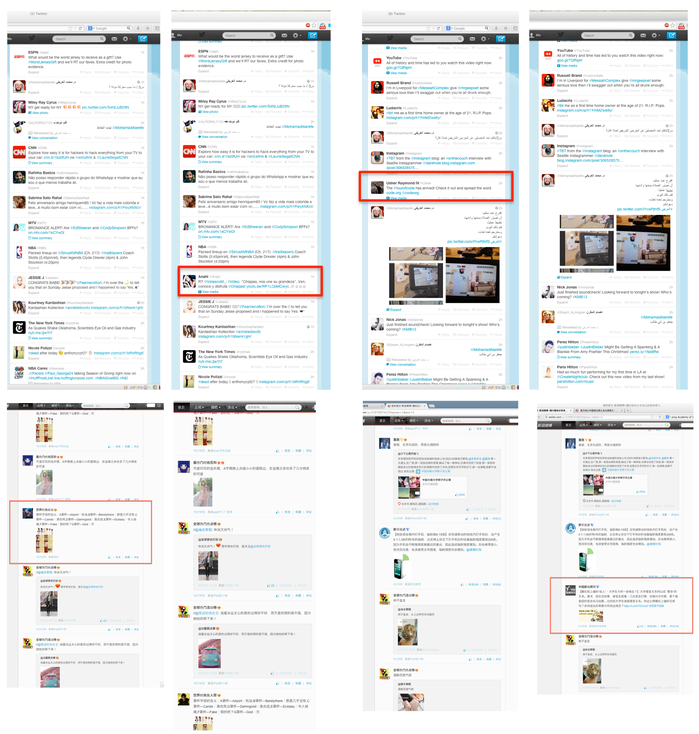
NOT caused by content negotiation, service localization, personalization, or randomization. These differences will not disappear over time, but relaxed consistency is volatile by nature.

Why Relax Consistency?

- A conscientious choice of the service provider
- CAP Theorem: a tradeoff
- Better user experience: latency and scalability over consistency

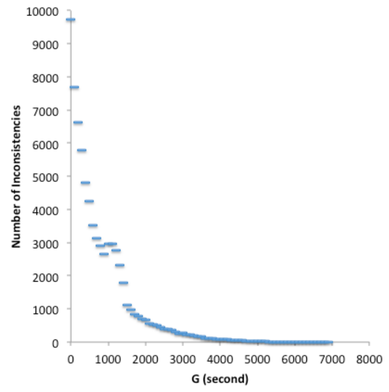
Myth

Inconsistency is barely noticeable, therefore harmless..



How Bad Is It? A Simulation.

- 6.27% contain observable conflicts
- Average inconsistency window is 823 seconds



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 resources are more prone to inconsistency

Key Takeaways

Web inconsistency may be more severe than we are aware of.

Possible Remedy

- Proactive approach: redundant crawl
- Compensatory approach: estimate inconsistency probability then label archival credibility

Los Alamos NATIONAL LABORATORY
hhmi | janelia Research Campus
UNM
VirginiaTech Invent the Future®