Nudging Users Away from Unsafe Content

Jian Liu and Sourav Bhattacharya

UNIVERSITY OF HELSINKI

- What signals can effectively nudge users away from unsafe content?
- Use machine learning to address time lag in crowdsourced risk signals.
- Identifying inappropriate content remains a challenge.

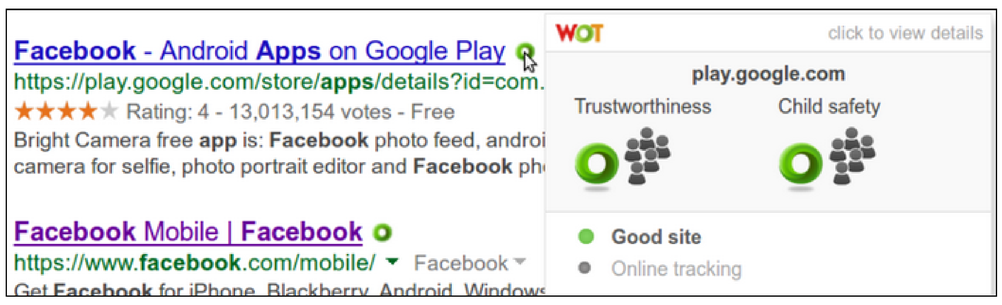
Problems

- Unsafe content (links, posts, apps) spreads virally
- Decisions difficult for users (lack of useful signals)
- Global ratings are not appropriate for all users
- Expert- and crowdsourced ratings suffer from time lags

Goals

- Find ways to overcome the time lag problem
- Find a new model for signaling inappropriateness
- Verify effectiveness of proposed solutions

WOT: An Example Crowdsourced Rating System



Web of Trust (WOT)

Overcoming Time Lag in WOT

Our approach uses page features to predict eventual rating

link	HTML.img.tags	JS.setTimeout	HOST.ns_ttl	...
vuupc.com	6	3	62320	...

Inferring likely WOT rating level from page features

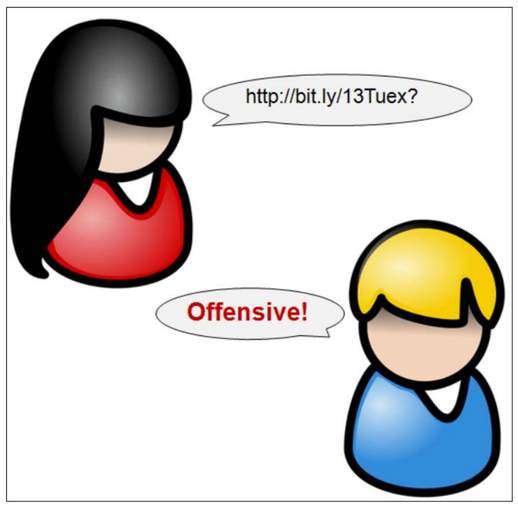
	Precision%	Recall%	F-score%
	90.6	93.2	91.9
	69.4	63.8	66.5
	89.9	87.0	88.5
Avg	89.2	89.3	89.2

Classification performance

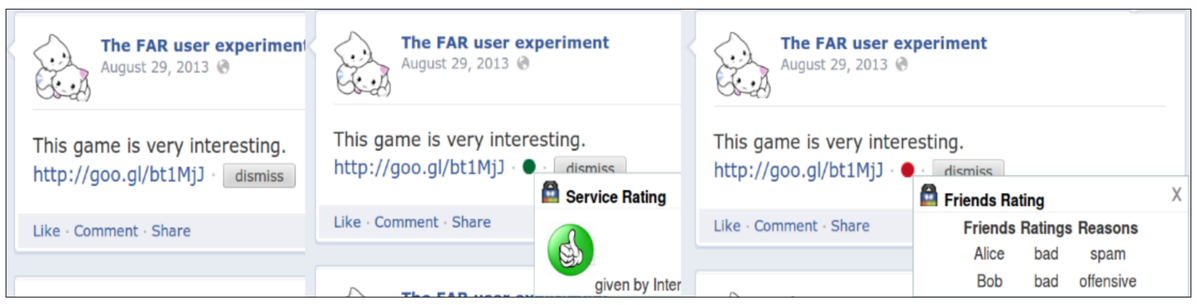
Groupsourcing sources information about unsafe content from users' social circles

Advantages

- Signaling inappropriateness
- Small time lag
- Trusted individuals
  - Better incentives
  - Visibility and traceability



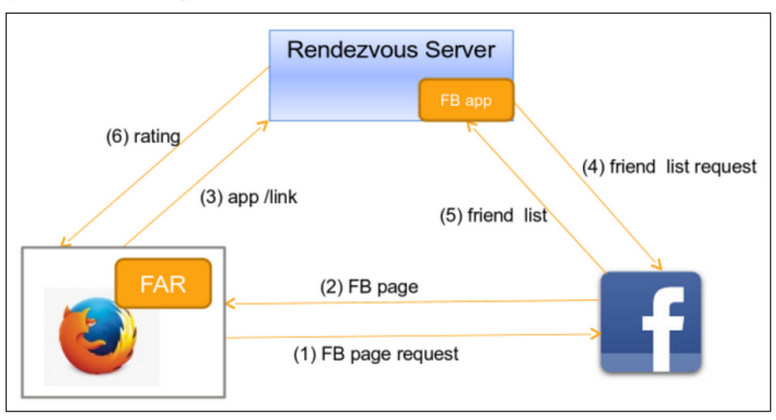
User Studies



Results

- Warning glyph (group- or centrally-sourced) significantly lower click-through rates (CTRs) for unsafe links
- No significant difference in CTRs between group- and centrally-sourced ratings
- 40% of red glyphs were clicked on

FAR allows savvy Facebook users to warn their friends about potentially unsafe content.



https://wiki.helsinki.fi/display/SecSys/FAR

Intel CRI for Secure Computing