

Re-Evaluating the Wisdom of Crowds in Assessing Web Security

Pern Hui Chia

Svein J. Knapskog

Centre for Quantifiable Quality of Service in Communication Systems (Q2S), NTNU

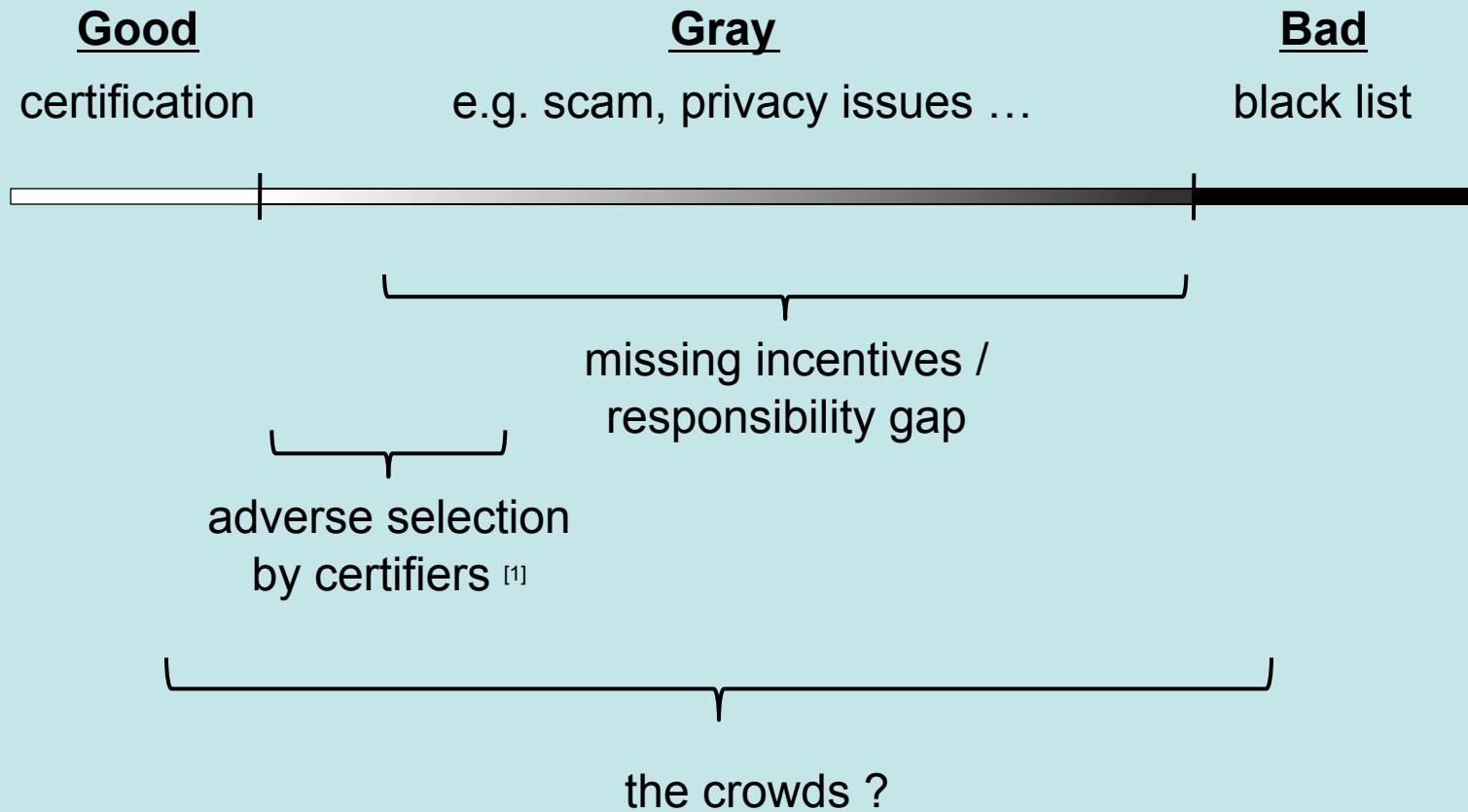
Financial Crypto 2011, St. Lucia, 28 Feb – 4 Mar 2011

Outline

- Background
- Data collection
- Evaluation
 - Reliability
 - Contribution pattern
 - Exploitability vs. objectivity
 - User concerns
- Discussion
- Conclusions

Background

- Web security remains challenging
 - online banking fraud loss in UK => £59.7m (2009), £24.9m (Jan-Jun 2010) ^[9]
 - >3m sites initiate drive-by downloads ^[4]
- Online adult industry: >\$97b (revenue, 2006) ^[6]
 - malware and script-based attacks
- Missing incentives to takedown?
 - phishing site (4 – 96 hours) ^[3]
 - mule recruitment (2 weeks) ^[3]
 - illegal pharmacies (2 months) ^[3]



The Wisdom of Crowds for Security

- For:
 - The many can be smarter than the few [5]
 - Advantage of scale
- Against / concern:
 - Reliability
 - Incentives
 - Gaming behavior
- PhishTank
 - Moore and Clayton (2008) [3]
 - Less comprehensive, less timely
 - Power-law contribution => susceptible to manipulation
 - Yet today, used by McAfee, Mozilla, Opera, Yahoo! ... [8]

Web Of Trust (WOT)

- Community-based site reputation system + browser add-on



- 30m sites rated
- 17m total users
 - 2m registered



- 4 evaluation aspects

- Trustworthiness: whether a site can be trusted, is safe, delivers what it promises
- Vendor Reliability
- Privacy
- Child Safety

- Each WOT rating (for each aspect) comes with a confidence level (rather than input count)

WOT : Default Signaling & Warning

Signaling : Trustworthiness

unless {
Vendor Reliability, or
Privacy }

has rating < 40% & conf. > 8%

* child-safety aspect ignored

Warning

if {
Trustworthiness,
Vendor Reliability, or
Privacy }

has rating < 40% & conf. > 8%

Screensavers for free

Lots of interactive **screensavers** to choose from... Thanks for the great **screensaver**, Tom. It sure looks stunning on my new laptop ... can't believe you give it away, that it's **free**, and ...
[example.com](#) - Similar pages

Free wallpapers, Screensavers, Flash & celebrity screensavers

Free wallpaper **screensavers**, **free** popular **screensavers**. Animal themes, cartoon themes, nature themes, star themes, ...
[example.net/screensavers](#) - Similar pages

The best free screensavers ever

We have all the **screensavers** you'll ever need. Animation, bikes, cars, cartoons, celebs, desktops ... **free** stuff and samples ...

WARNING!

CATEGORIES SCREENSAVERS WALLPAPERS PLAY GAMES

This site has a poor reputation.
screensavers.com

[View rating details and comments](#)

Trustworthiness	Poor
Vendor reliability	Very poor
Privacy	Very poor
Child safety	Very poor

[This site is safe - I want to rate it](#) [Ignore warning and go to the site](#)

Data Collection

1

Evaluation on 20k domains (from Alexa top-million sites)

- WOT
- McAfee SiteAdvisor (SA)
- Norton Safe Web (SW)
- Google Safe Browsing Diagnostic Page (SBDP)

2

WOT

Count of ratings & comments of 50k random users

3

WOT

485k random comments +
Aggregate ratings on 412k unique domains

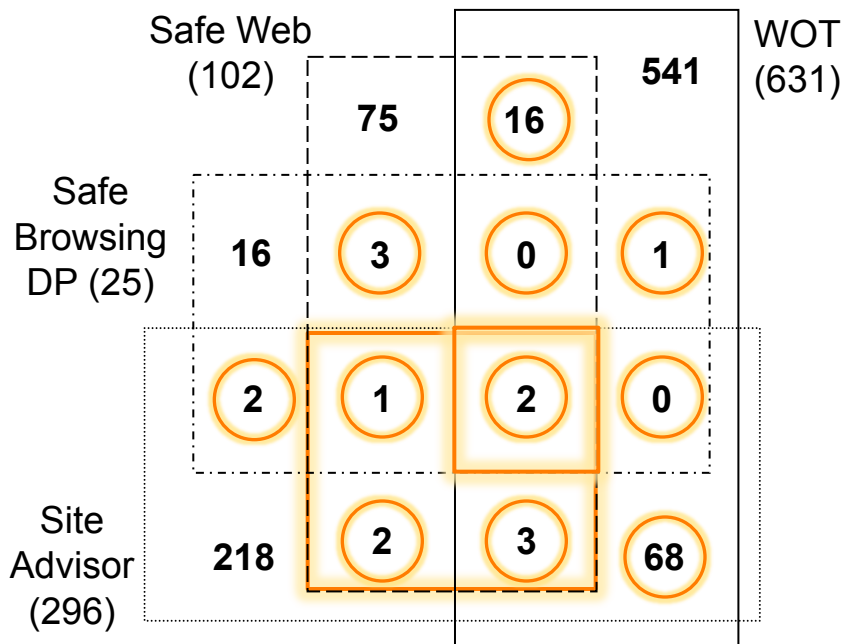
Evaluation & Results

Coverage for general sites

	Cov. (%)	Outcomes (%)		
		Bad	Caution	Good
WOT	51.23	3.16	2.15	45.93
SA	87.84	1.48	0.47	85.90
SBDP	55.65	0.13	1.63	53.90
SW	68.09	0.51	0.38	67.21

- Align outcomes into “Good”, “Caution”, “Bad”, “Unknown”
- WOT: Lower coverage expected for user-based system?
 - Increases considering sites from US, CA, EU, NO, CH only
- WOT: 3.16% bad sites => broader scope?
- SW, SBDP: too optimistic?

Divergent evaluation outcomes



948 total 'bad' sites

2 receive the same verdict from all

8 on common blacklist of SA and SW

98 (10%) regarded as 'bad' by >1 services

=> Different scope and method,
Lack of sharing

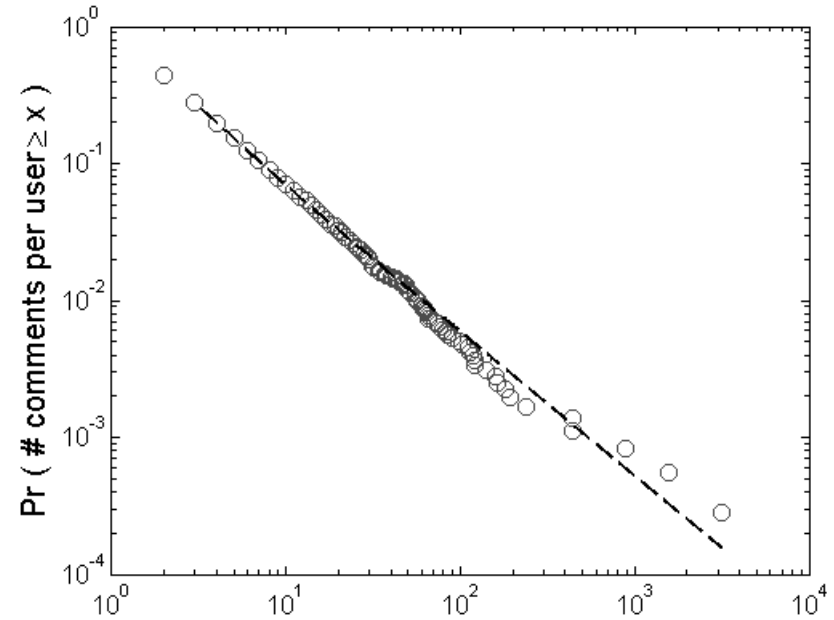
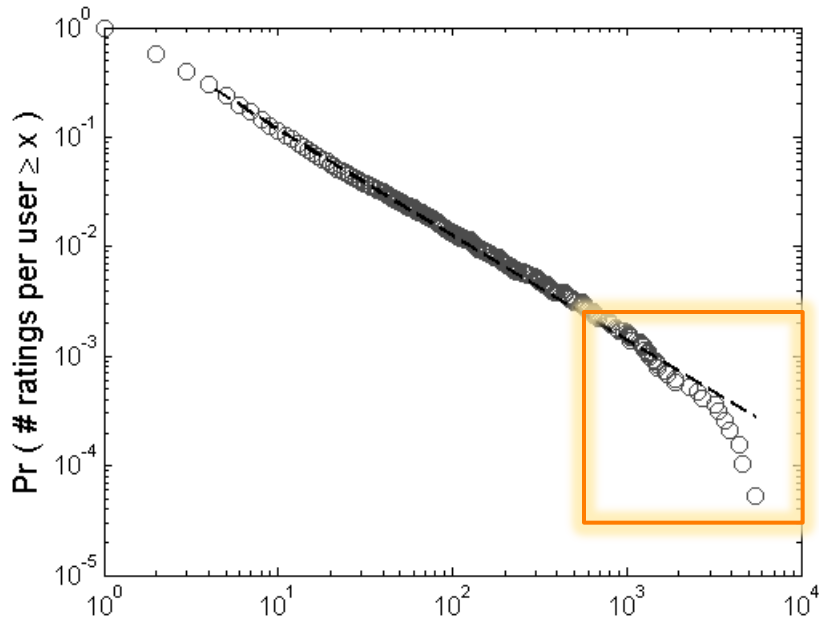
Recall, Precision in identifying 'bad' sites

	Conservative consensus (%)					
	R	P	FS	$F_{n,g}$	$F_{n,u}$	$F_{n,c}$
WOT	22.1	14.3	17.3	22.6	49.4	5.9
SA	10.7	26.4	15.2	69.7	15.6	4.0
SW	3.1	26.5	5.5	71.6	23.6	1.7
SBDP	1.0	36.0	1.9	47.6	46.2	5.2
WOT [credible]	17.2	17.8	17.5	-	-	-
SA [auto]	8.3	3.4	4.8	68.6	20.7	2.5
SW [auto]	3.5	10.8	5.3	66.1	26.9	3.5
SBDP [auto]	2.1	32.0	3.9	43.6	44.6	9.7

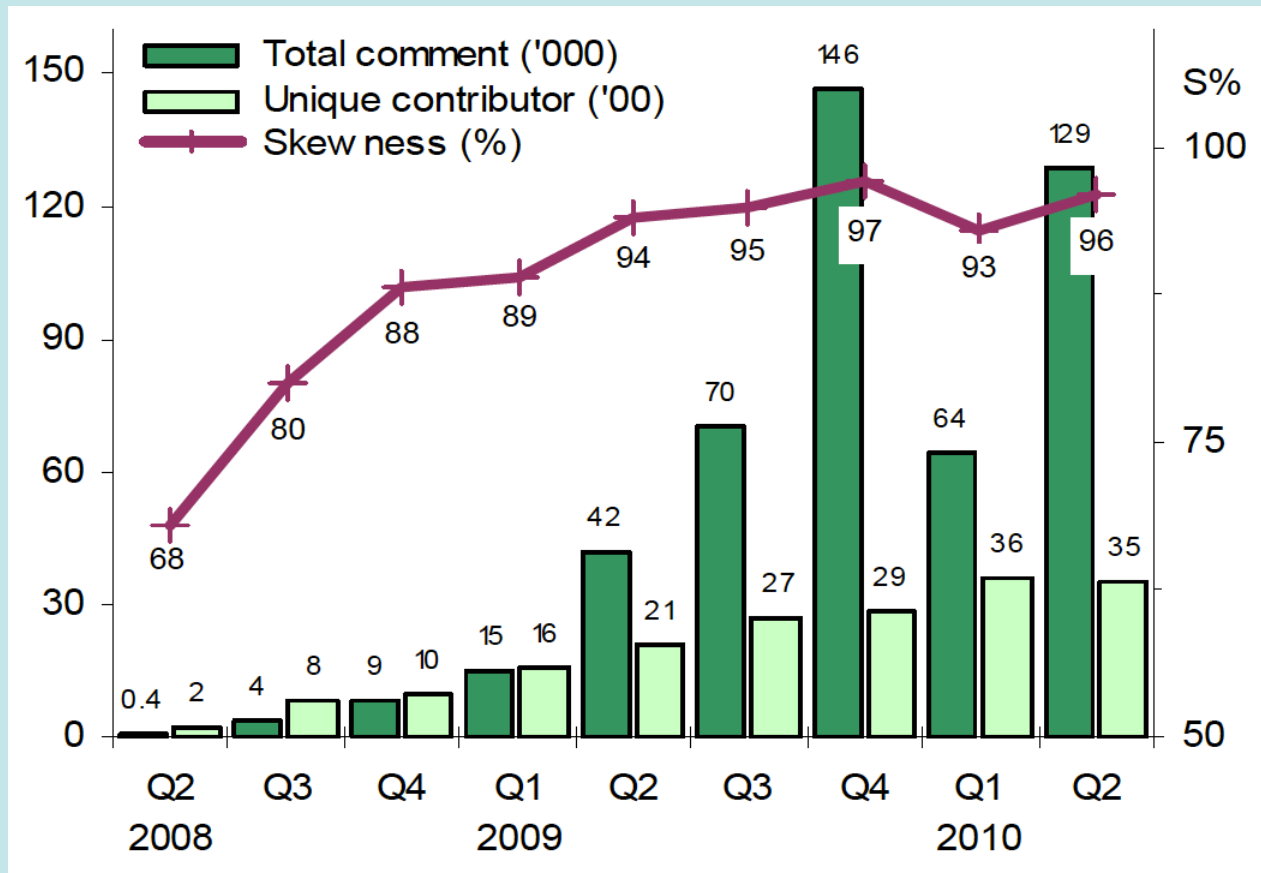
- WOT: higher Recall, lower Precision (broader scope?), higher F-Score
 - Results unchanged considering cases with credible warning
- WOT: False Negatives mostly labeled as 'unknown'
 - SA, SW mostly wrongly classify FN as 'good'
- Compare among automated services only => SA, SW, SBDP all have low F-Score

Evaluation: Dataset-II

Contribution Pattern



- Skewed contribution ratios
 - Comment : Power Law ($\alpha=2.05$, $x_{\min}=3$)
 - Rating : not Power Law (curve-in for highly active users)



- No. of unique contributors & no. of comments increase
- But, comment contribution distr. also more skewed with time
 - Mostly due to mass rating tool given to highly active users
 - Introduced in Sep 2008: privilege for both gold and platinum users
 - New privilege given to only Platinum users from Dec 2009

- Risk mitigation measures by WOT
 - Inputs weighted based on contributor's reliability, not activity level
 - Manipulation may still be possible, but comes with a high cost
 - Automated detection of unusual rating behavior

Exploitability vs. Objectivity

- Let 'conflict' = when a positive comment is given to a site with poor rating, or vice versa

Probable factors:

1. Comment out of rating scope
 2. Exploitation
 3. Opinion difference
- } subjectivity / non-verifiability

- Indeed, comment categories on
 - user experience : %-conflict >5%
 - browser exploits, phishing sites, adult content : low %-conflict
- Low %-conflict \neq a small number of unique contributors
- WOT: 90% comments in categories with %-conflict < 5%
=> verifiable

- Filtered for unique comments & processed words
- Concerns not limited to phishing and malware
 - Popular nouns used:
 - Spam
 - Scam
 - Information (used with 'personal' and 'sensitive')
 - Pharmacy
 - Phishing
 - Malware, virus and Trojan

Discussion

Discussion:

Limitations of Our Study

- We did not evaluate against
 - Malicious sites in the long tail of web popularity
 - Short-lived malicious sites (i.e., timeliness of outcomes)
- But, we note that WOT does handle above concerns using inputs from trusted black-lists (e.g., PhishTank, SpamCop, LegitScript)

Discussion:

Potential lessons from WOT

- Risk mitigation measures:
 - Inputs weighted based on the contributor's reliability
 - Bayesian updating of individual's reliability
 - Automated detection of unusual rating behavior
 - Virtual community: discussion forum, mass rating tool user must be contactable, etc
- Ease to contribute / efficiency:
 - Add-on makes rating easy
 - Sub-domains inherit the reputation of parent domain
 - Weight decays with time, but replenishes as user re-visits a site

Potential Weaknesses of WOT

- Skewed participation ratio
 - Mass-rating tool should be handled with care
 - Important to diversify the sources of contribution
 - But, a necessary phase of a community-based system? [11]
- Details of rating and reliability computation are hidden
 - May be hard to believe the ‘false positives’
 - But, aren’t the details of many other practical algorithms hidden?
- Subjective evaluation factors
 - Distinction of objective and subjective criteria can help

Conclusions

- WOT
 - More comprehensive in identifying ‘bad’ domains
 - Skewed contribution ratios. But,
 - Built-in measures to mitigate risks of exploitation
 - Majority inputs based on verifiable factors
- User concerns not limited to malware / phishing
 - Potential for user-based systems to complement conventional methods
- We do not play down the role of automated blacklists, but our study shows that
 - **wisdom of crowds for security can work** with careful design

Reference

1. Edelman, B., “Adverse selection in online “trust” certifications and search results,” *Electronic Commerce Research and Applications*, 2010.
2. Moore, T., and Clayton, R., “Evaluating the Wisdom of Crowds in Assessing Phishing Websites,” *FC 2008*.
3. Moore, T., and Clayton, R. The Impact of Incentives on Notice and Takedown. In *Managing Information Risk and the Economics of Security*, ed. Johnson, M.E., 2008.
4. Provos, N., Mavrommatis, P., Rajab, M.A., and Monrose, F. All your iFRAMEs point to Us, In *Proc. USENIX Security 2008*.
5. Surowiecki, J. The wisdom of crowds. Anchor Books, 2005.
6. Wondracek, G., Holz, T., Platzer, C., Kirda, E., and Kruegel, C. Is the Internet for Porn? An Insight into the Online Adult Industry, In *Proc. WEIS 2010*.
7. Zhuge, J., Holz, T., Song, C., Guo, J., Han, X., and Zou, W. Studying Malicious Websites and the Underground Economy on the Chinese Web. In *Proc. WEIS 2008*.
8. Friends of PhishTank. <http://www.phishtank.com/friends.php>
9. The UK Card Association. New Card and Banking Fraud Figures.
www.theukcardsassociation.org.uk/media_centre/press_releases_new/-/page/922/
www.theukcardsassociation.org.uk/media_centre/press_releases_new/-/page/1037/
10. WOT statistics: <http://www.mywot.com/en/community/statistics>
11. Kittur, A., Chi, E. H., Pendleton, B. A., Suh, B., and Mytkowicz, T. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Alt.CHI 2007*.

Thank you. Question?

Pern Hui Chia
chia@q2s.ntnu.no

Svein J. Knapskog
knapskog@q2s.ntnu.no

	WOT	SiteAdvisor	Safe Browsing DP	Safe Web
Good	Tr \geq 60, and no credible warning in Vr or Pr	Green: very low or no risk	Site not currently listed as suspicious, and Google has visited it in the past 90 days.	Safe
Caution	60 > Tr \geq 40, and no credible warning in Vr or Pr	Yellow: minor risk	Site not currently listed as suspicious, but part of the site was listed for suspicious activity in the past 90 days.	Caution
Bad	Tr < 40, or there is a credible warning in Vr or Pr	Red: serious risk	Site is listed as suspicious.	Warning
Unknown	No Tr rating, and no credible warning in Vr or Pr	Gray: not rated	Site not listed as suspicious, and Google has not visited it the past 90 days.	Untested

Aligning the evaluation outcomes

	Cov.	Outcomes (%)		
	(%)	Bad	Caution	Good
WOT	51.23 ^a	3.16	2.15	45.93
SA	87.84	1.48	0.47	85.90
SBDP	55.65 ^b	0.13	1.63	53.90
SW	68.09	0.51	0.38	67.21

Coverage and percentage of evaluation outcomes

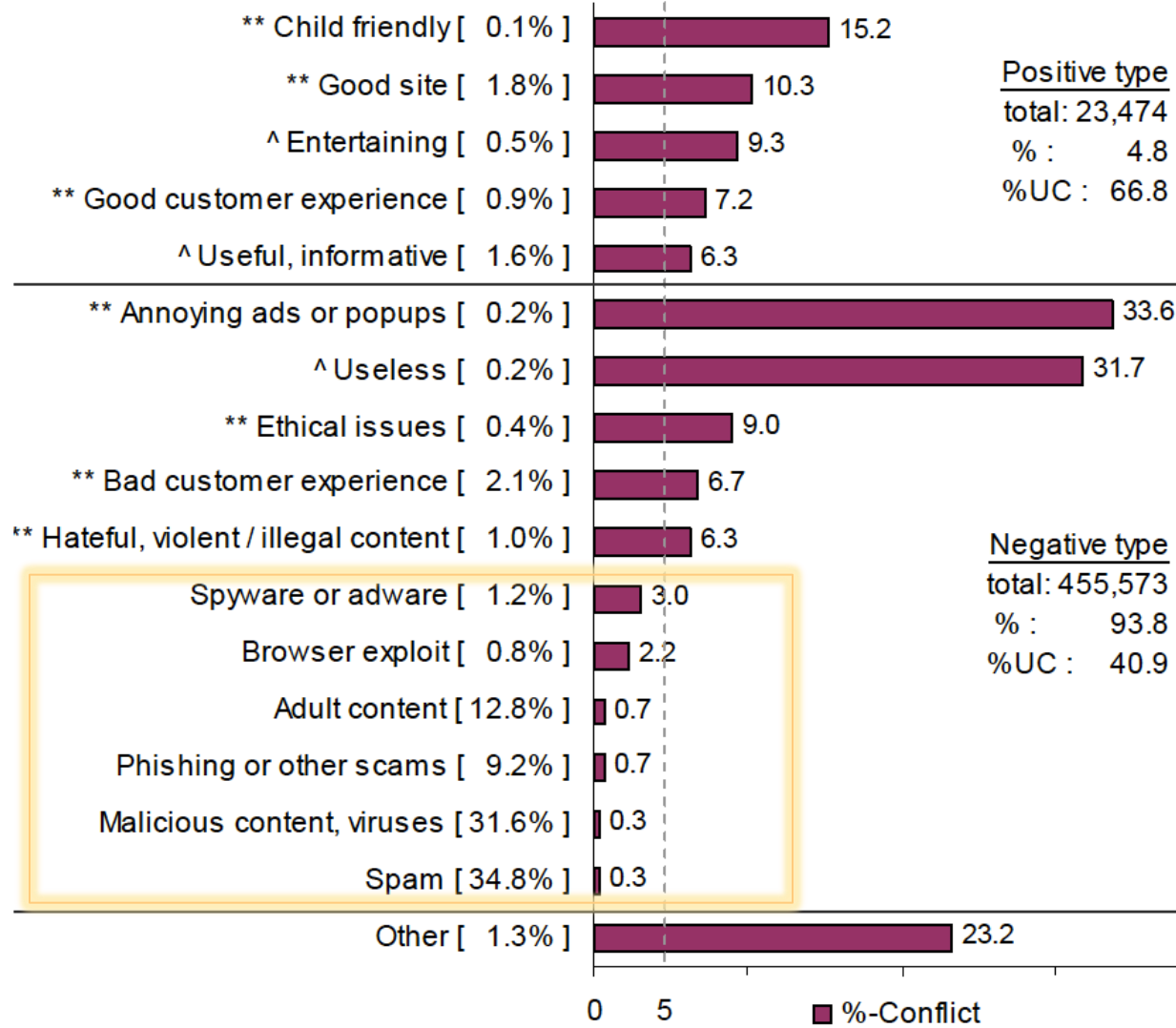
- a. based on default risk signaling strategy of WOT and not including sites that has only the child-safety rating
- b. we regard sites not currently blacklisted and have not been visited by Google in past 90 days as 'not tested'

	Optimistic consensus (%)						Conservative consensus (%)					
	R	P	FS	$F_{n,g}$	$F_{n,u}$	$F_{n,c}$	R	P	FS	$F_{n,g}$	$F_{n,u}$	$F_{n,c}$
WOT	15.3	1.7	3.1	11.1	72.2	1.4	22.1	14.3	17.3	22.6	49.4	5.9
SA	8.3	3.4	4.8	57.5	27.5	6.7	10.7	26.4	15.2	69.7	15.6	4.0
SW	4.1	8.8	5.6	59.0	34.2	2.7	3.1	26.5	5.5	71.6	23.6	1.7
SBDP	2.5	16.0	4.3	40.0	55.6	1.9	1.0	36.0	1.9	47.6	46.2	5.2
WOT [credible]	13.9	2.5	4.3	-	-	-	17.2	17.8	17.5	-	-	-
SA [auto]	10.0	2.0	3.4	68.3	18.3	3.3	8.3	3.4	4.8	68.6	20.7	2.5
SW [auto]	2.9	4.9	3.6	65.7	28.0	3.4	3.5	10.8	5.3	66.1	26.9	3.5
SBDP [auto]	3.3	16.0	5.4	42.3	51.2	3.3	2.1	32.0	3.9	43.6	44.6	9.7

Recall, Precision and F-Score in Optimistic and Conservative consensus

Finding of this service	Findings of other services				
	Bad w/o any good	Mixed of good & bad	Caution only	Good w/o any bad	All unknown
Bad	T_p [T_p]	$F_{p,m}$ [T_p]	$F_{p,c}$ [$F_{p,c}$]	$F_{p,g}$ [$F_{p,g}$]	$F_{p,u}$ [$F_{p,u}$]
Caution	$F_{n,c}$ [$F_{n,c}$]	- [$F_{n,c}$]	-	-	-
Good	$F_{n,g}$ [$F_{n,g}$]	- [$F_{n,g}$]	-	-	-
Unknown	$F_{n,u}$ [$F_{n,u}$]	- [$F_{n,u}$]	-	-	-

True positives, false positives and false negatives



- 90% comments in categories with %-conflict < 5%
- Low %-conflict ≠ small number of unique contributors