CyLab Privacy Interest Group
2006 Privacy Policy Trends Report
Final Report

Lorrie Faith Cranor

20 February 2006

Carnegie Mellon

CMU **U**sable **P**rivacy and **S**ecurity Laboratory

http://cups.cs.cmu.edu/

Report Outline

- Executive summary
- Introduction
- Privacy policies of Popular and Random web sites
- Focus on financial industry
- Platform for Privacy Preferences
- Discussion

Popular and Random Websites

- Used list of 30,000 most clicked on domains from AOL search engine to create site lists
- Examined only .com domains that were not adult sites or kids sites (to be comparable with previous FTC studies)
- Popular list is top 100 domains on list that met above criteria
- Random list is 75 websites randomly selected from top 12,000 sites on that list, excluding sites not meeting above criteria (12,000 site sample frame selected based on growth of Internet since previous studies)

Differences Between Random Sample Groups

	1999	2000	2001	2006
Number of sites surveyed	***n* = 286**	***n* = 281**	***n* = 223**	***n* = 100**
Privacy policy	**48.3%**	**65.8%**	**76.7%**	**88%**
Provides notice about what personal information is collected	**49.7%**	**71.2%**	**73.5%**	**100%**
Provides notice about disclosure to third parties	**40.6%**	**N/A**	**N/A**	**83%**
Provides access	**27.6%**	**21.4%**	**N/A**	**94%**

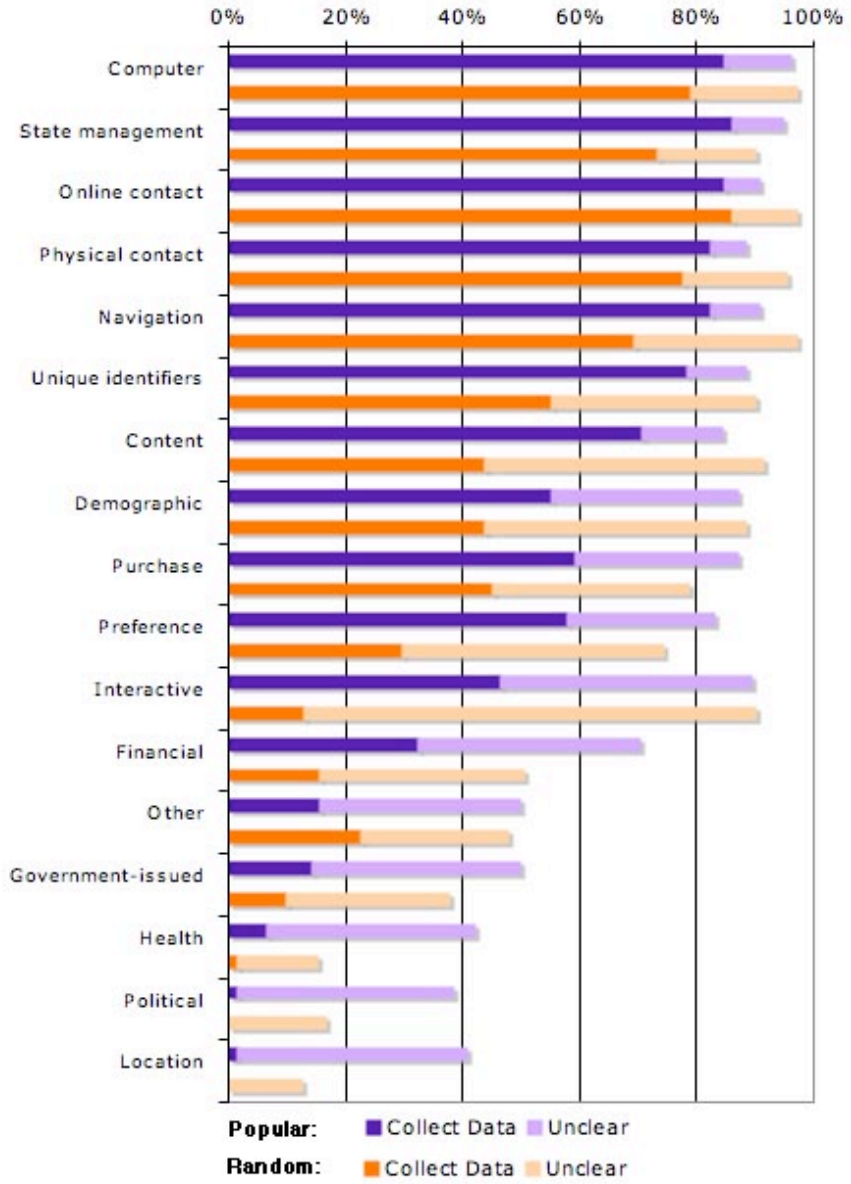
Differences Between Popular Sample Groups

	1998	1999	2000	2001	2006
Number of sites surveyed	***n* = 105**	***n* = 91**	***n* = 87**	***n* = 71**	***n* = 75**
Privacy policy	**44.8%**	**84.6%**	**96.6%**	**98.6%**	**96%**
Provides notice about what personal information is collected	**N/A**	**73.6%**	**90.8%**	**95.8%**	**100%**
Provides notice about disclosure to third parties	**58.1%**	**71.4%**	**N/A**	**N/A**	**80%**
Provides access	**26.7%**	**41.8%**	**49.4%**	**N/A**	**95%**

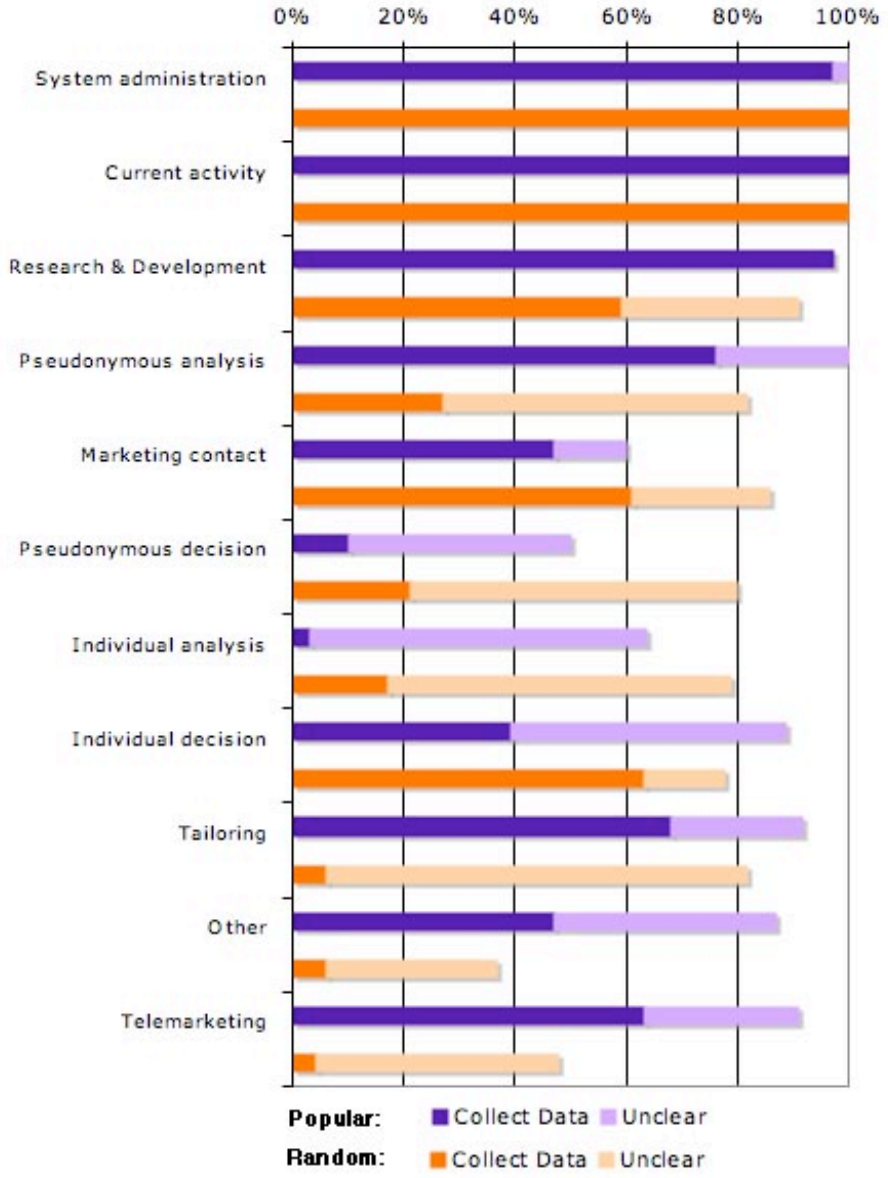
Privacy Bird evaluation

- **Low:** Sites must not collect health info and share it with other companies or use it for analysis, marketing, or to make decisions that may affect what content or ads the user sees. Sites must not engage in marketing without opt-out. **86%** of popular, **70%** of random
- **Medium:** Same as low, plus sites must not share PII, financial info, or purchase info with other companies; and sites that collect personally identified data must provide access provisions. **58%** of popular, **48%** of random
- **High:** Same as medium, plus sites must not share any personal info or use it to determine the user's habits, interests, or other characteristics; and sites may not contact users for marketing or use financial or purchase info for analysis, marketing, or to make decisions that may affect what content or ads the user sees. **31%** of popular, **17%** of random
- A site is classified as not sharing data if it shares data only under an opt-in policy or only with agents that use it only to complete the transaction for which it was provided or with delivery companies.
- Data from the following P3P categories are considered PII: physical contact info, online contact info, and government issued identifiers.

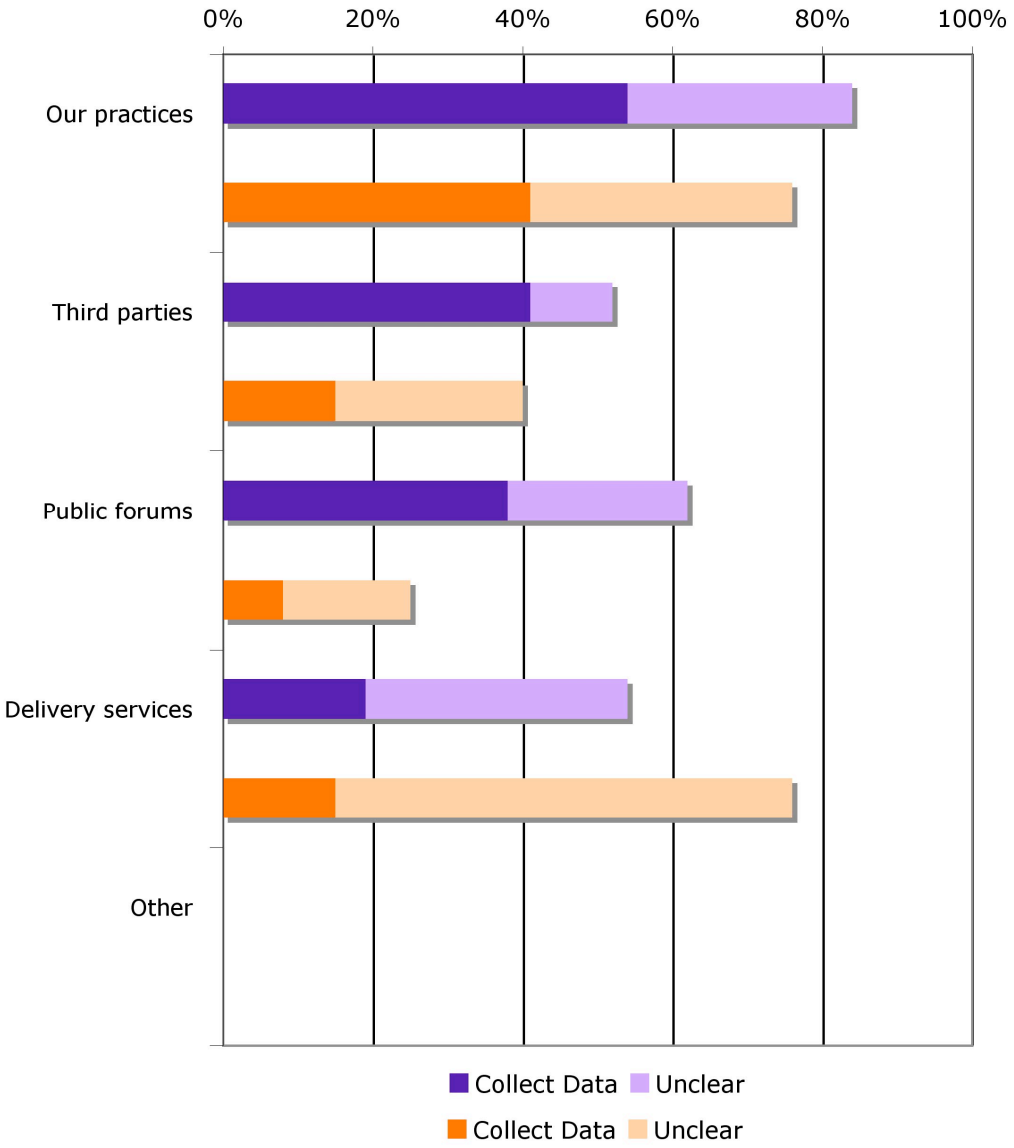
Data collected



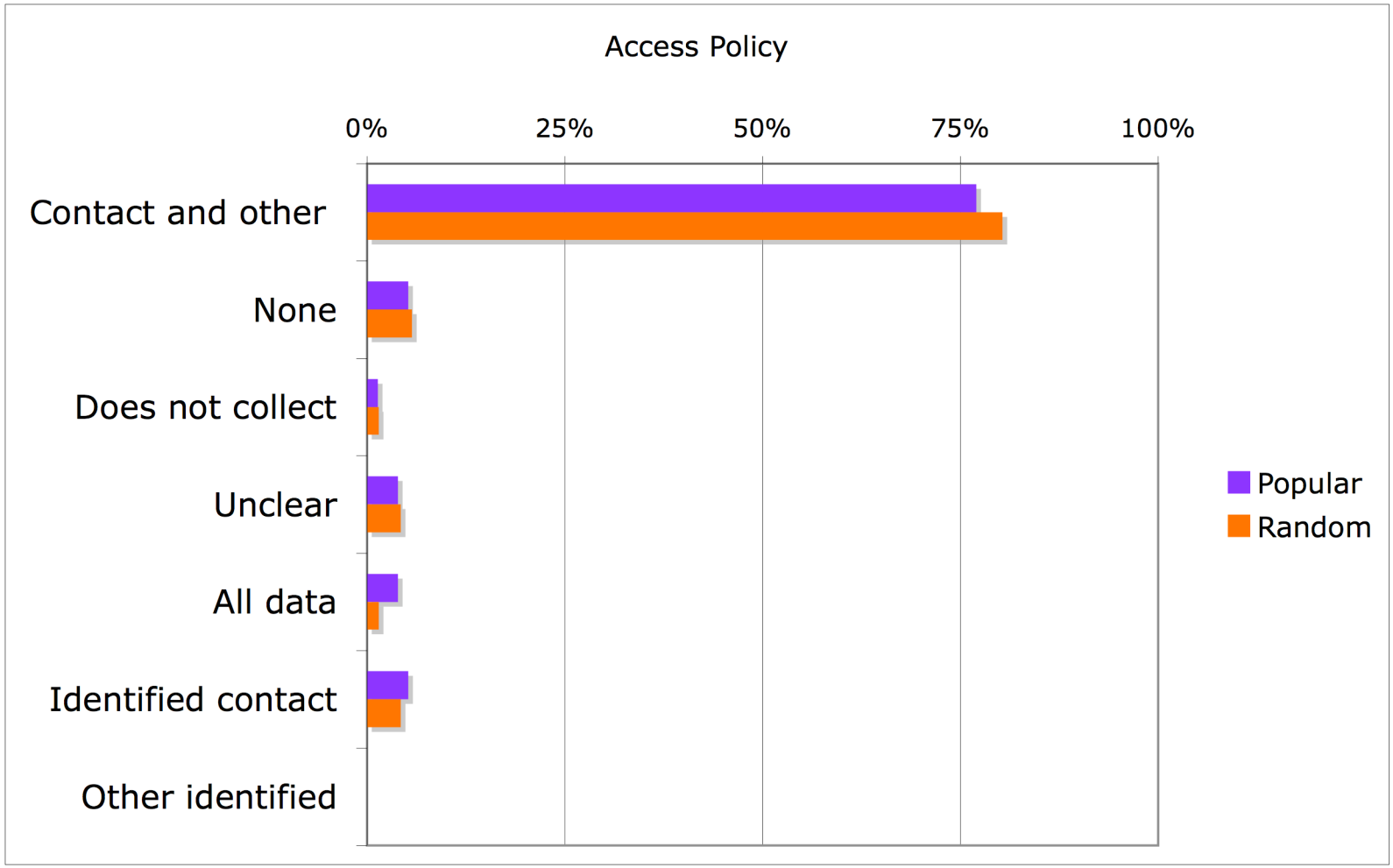
Data use


0%
20%
40%
60%
80%
100%
System administration
Current activity
Research & Development
Pseudonymous analysis
Marketing contact
Pseudonymous decision
Individual analysis
Individual decision
Tailoring
Other
Telemarketing
Popular:
Collect Data
Unclear
Random:
Collect Data
Unclear

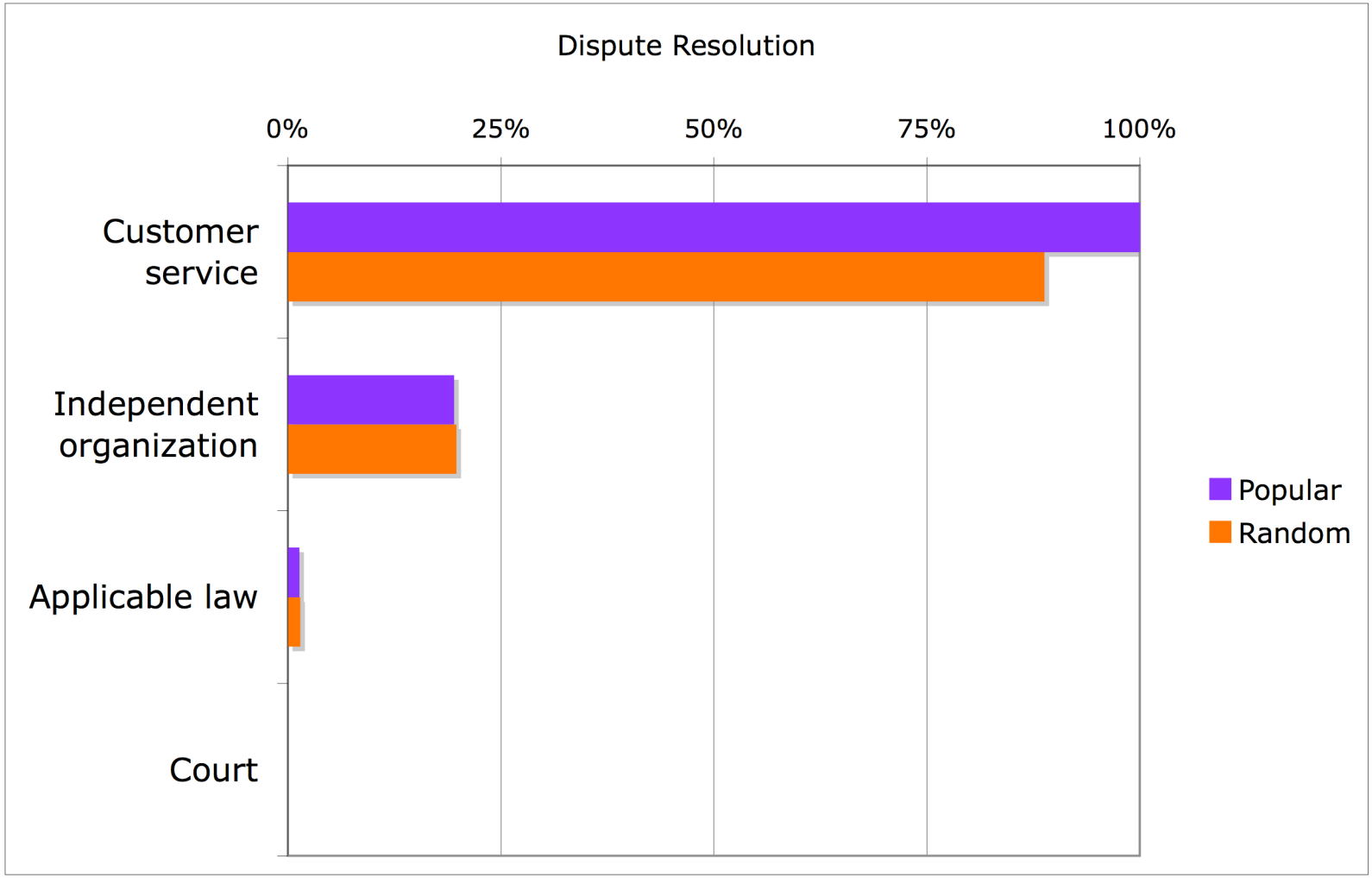
Data recipients



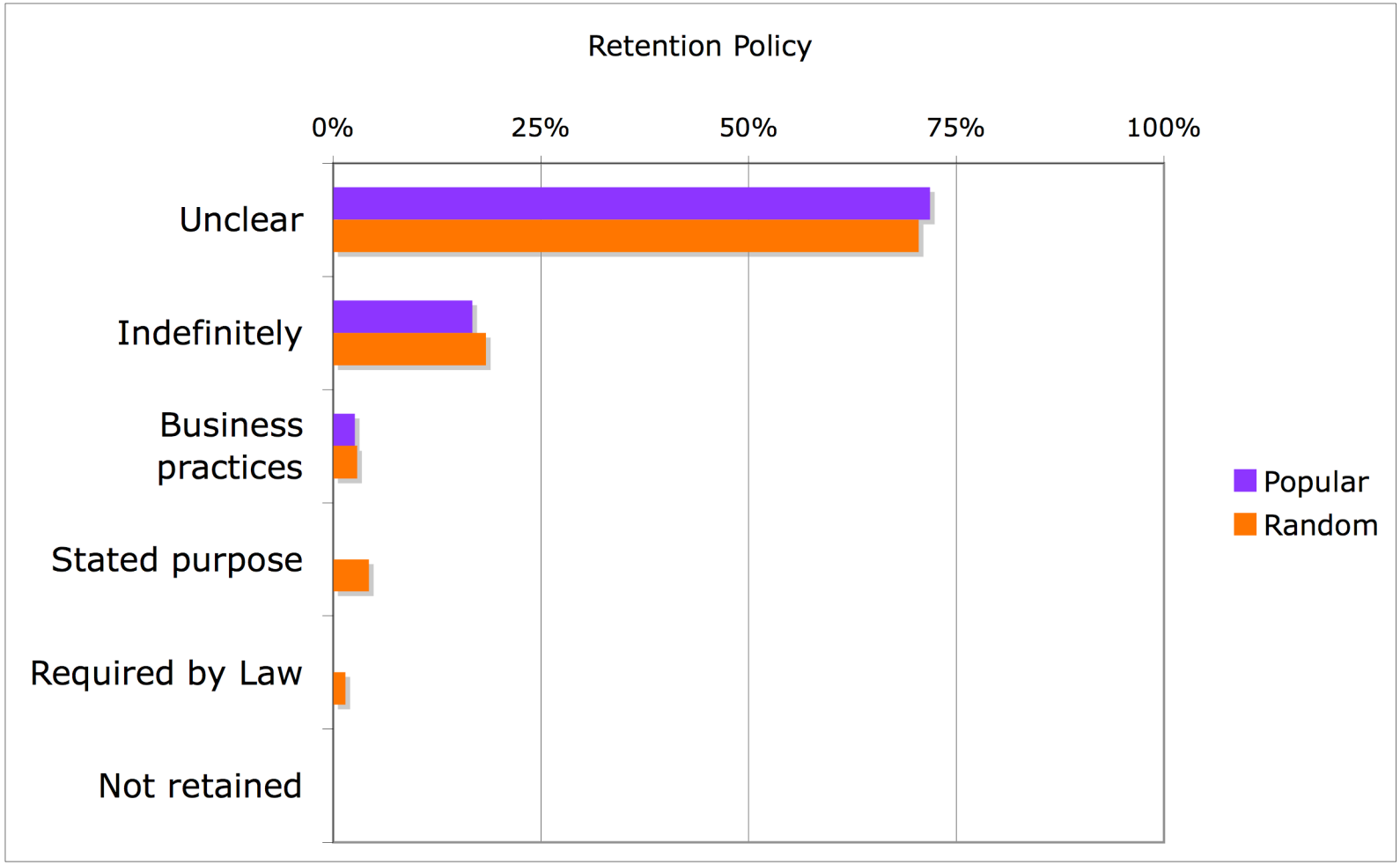
Access provisions



Dispute resolution



Data retention policy



Kincaid readability score

- Popular - 12.4 (standard deviation = 1.8)
- Random - 12.5 (standard deviation = 2.3)
- White house press release - 4.1
- New York Times article - 6.2
  - Readability Info, "Readability Grades," http://readability.info/info.shtml, Accessed 6 November 2006.

Focus on financial industry

- The readability and clarity of policies for financial institutions has improved
- Banks minimally comply with GLB in terms of affiliate sharing
- GLB had little impact on the third party sharing choices available to consumers
- *Details in report*

Platform for Privacy Preferences Project (P3P)

- Developed by the World Wide Web Consortium (W3C) **http://www.w3.org/p3p/**
  - Final P3P1.0 Recommendation issued 16 April 2002
- Standard machine-readable format for web site privacy policies
  - Can be deployed using existing web servers
- Enables the development of tools (built into browsers or separate applications) that
  - Summarize privacy policies
  - Compare policies with user preferences
  - Alert and advise users
- P3P support built into IE6 and Netscape 7

Our P3P policy database

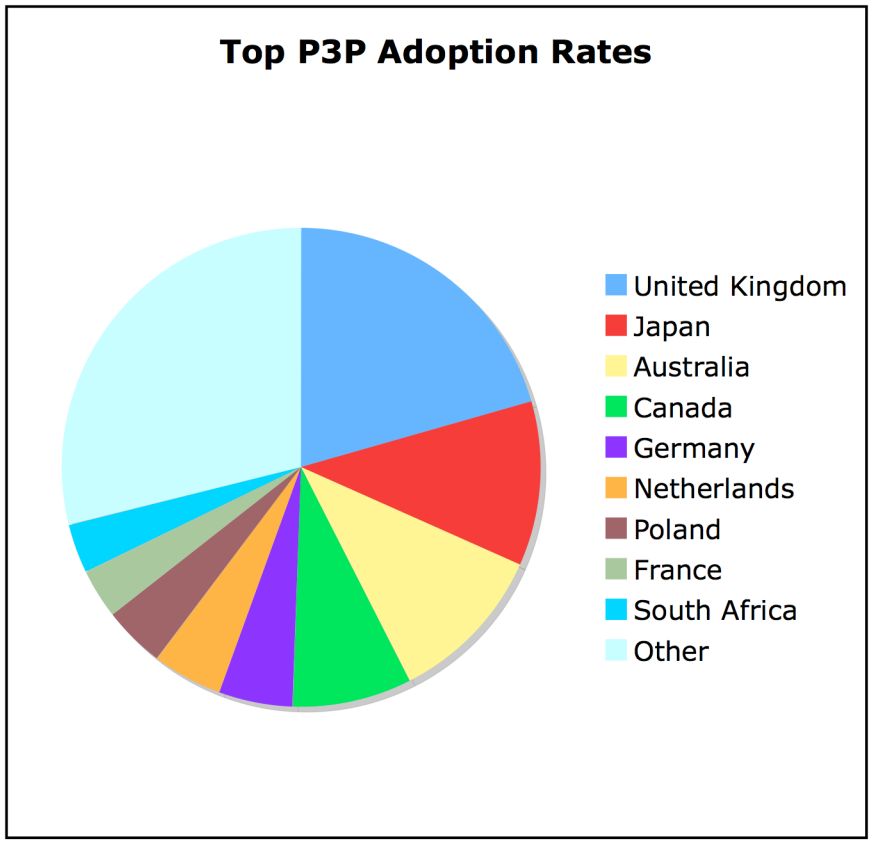
- Our P3P policy database is based on our Privacy Finder search engine cache
- Whenever someone runs a search, the 10 displayed results get checked for P3P policies and added to our cache
- We seeded the cache by running 20,000 queries provided by AOL and by checking for P3P policies at 30,000 most clicked on domains
- We have checked 113,880 Web sites for P3P policies and collected 11,843 policies, including 3,846 unique policies
- We revisit sites with P3P policies about once a day to see if the policy has changed
- We revisit other sites about once a month to see if they have added a P3P policy

Longitudinal Trends

- Sharpest increase in government sites, probably due to the E-Government Act (1,465%)
- Other large changes in sites targeted to children (106%) and news sites (43.95%)

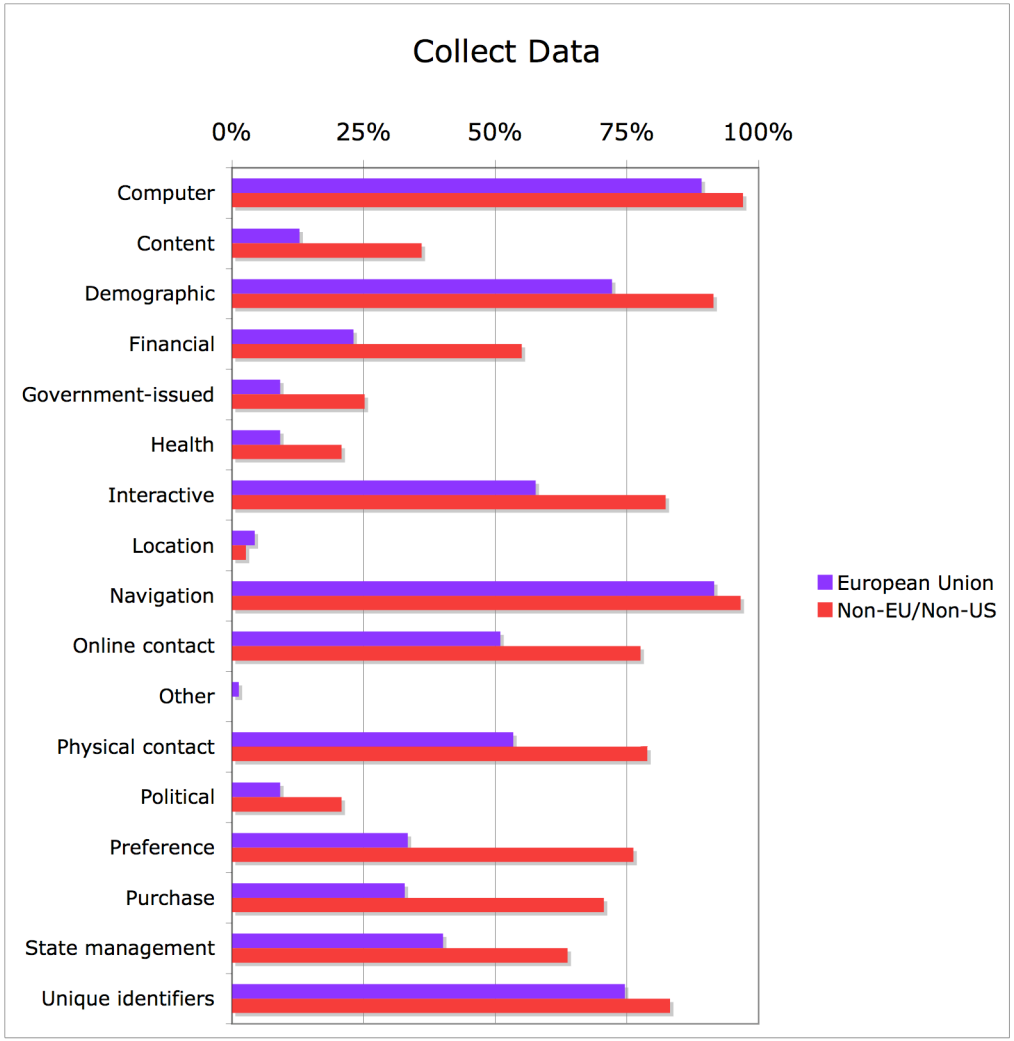
	# in list	Sites reached in 2003	P3P-enabled in 2003	Sites reached in 2006	P3P-enabled in 2006	% change
PFF Random	302	286	12.23%	282	10.99%	-10.14%
PFF Most Popular	85	84	30.95%	84	25.00%	-19.22%
PFF Refined Random	209	195	14.87%	195	12.82%	-13.79%
Key Measures	500	486	23.46%	474	23.63%	+0.72%
Netscore Top 500	500	488	22.95%	474	23.84%	+3.88%
Alexa	500	495	18.59%	470	18.51%	-0.43%
FirstGov	344	338	2.07%	321	32.40%	+1465.22%
Froogle	1017	1010	13.17%	964	12.55%	-4.71%
News	2429	2398	9.42%	2286	13.56%	+43.95%
Yahooligans!	900	868	3.00%	841	6.18%	+106.00%
Total	**5856**	**5739**	**10.25%**	**5414**	**13.59%**	**+32.59%**

Geographic distribution

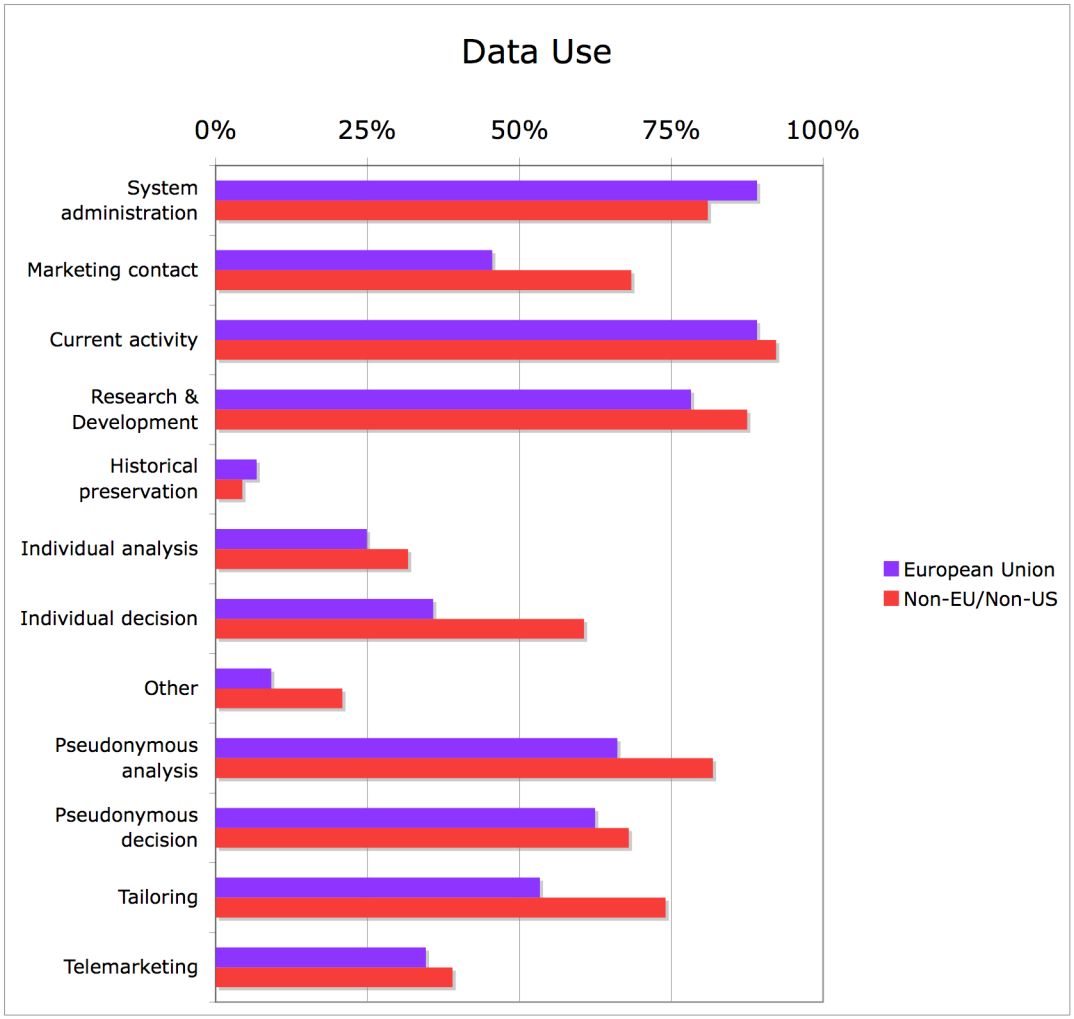


- 437 of 8,661 P3P policies in database have non-US country-specific TLD
  - UK 91
  - Japan 49
  - Australia 48
  - Canada 35
  - Germany 22
  - Netherlands 22
  - Poland 18
  - France 15
  - South Africa 14
  - …
- 45% of non-US TLDs in EU

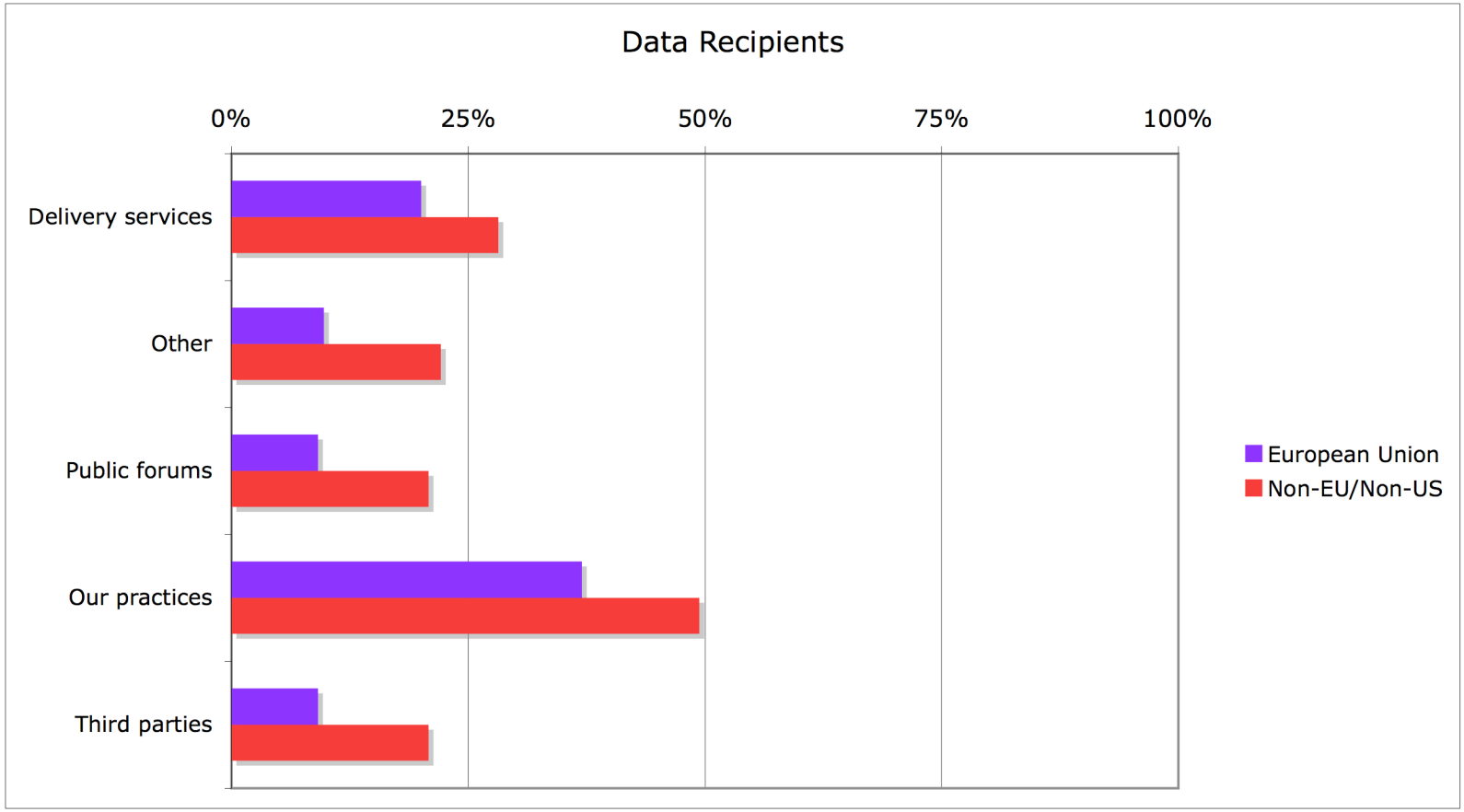
Data collected (EU v. Non-EU/Non-US)



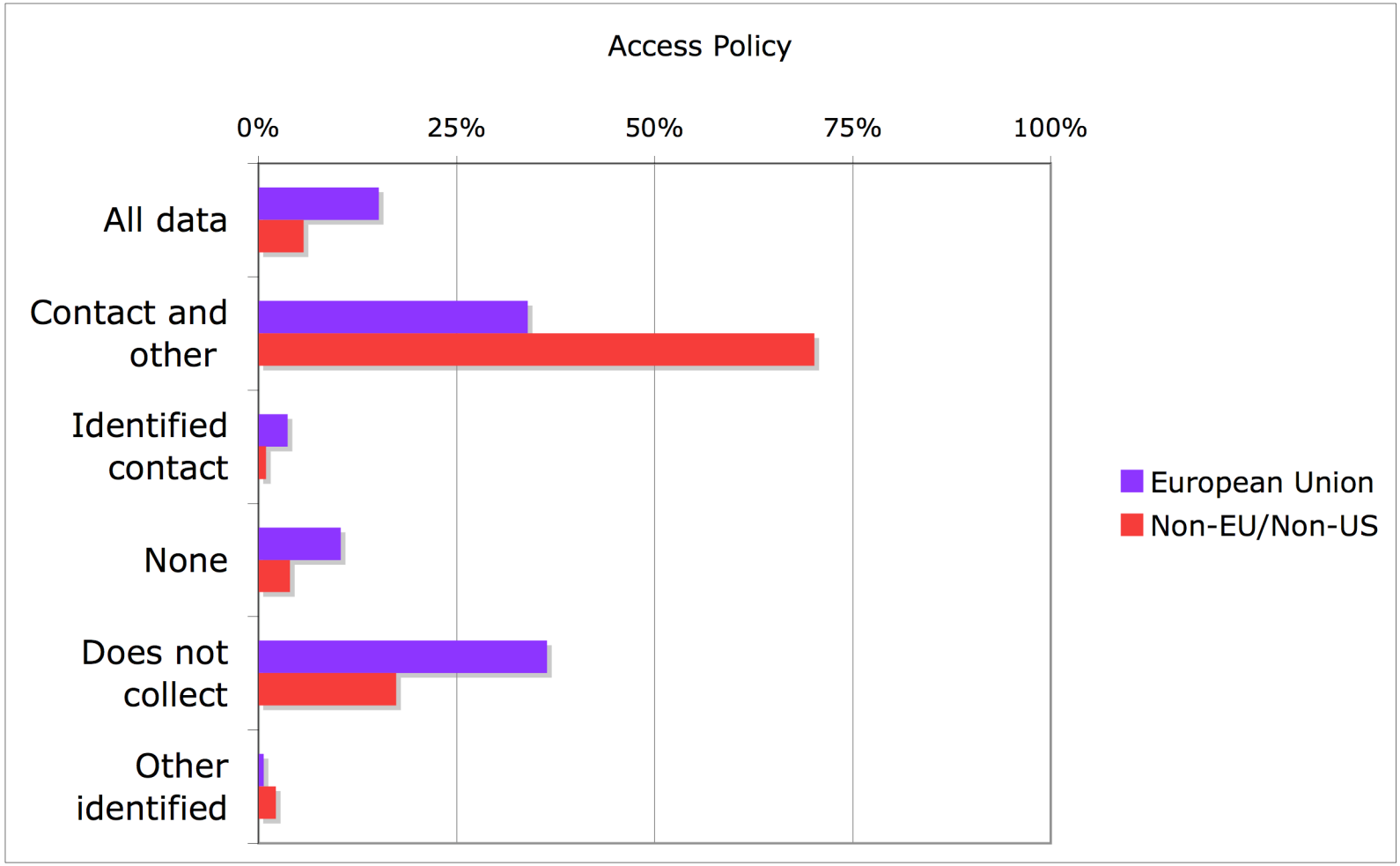
Data use (EU v. Non-EU/Non-US)



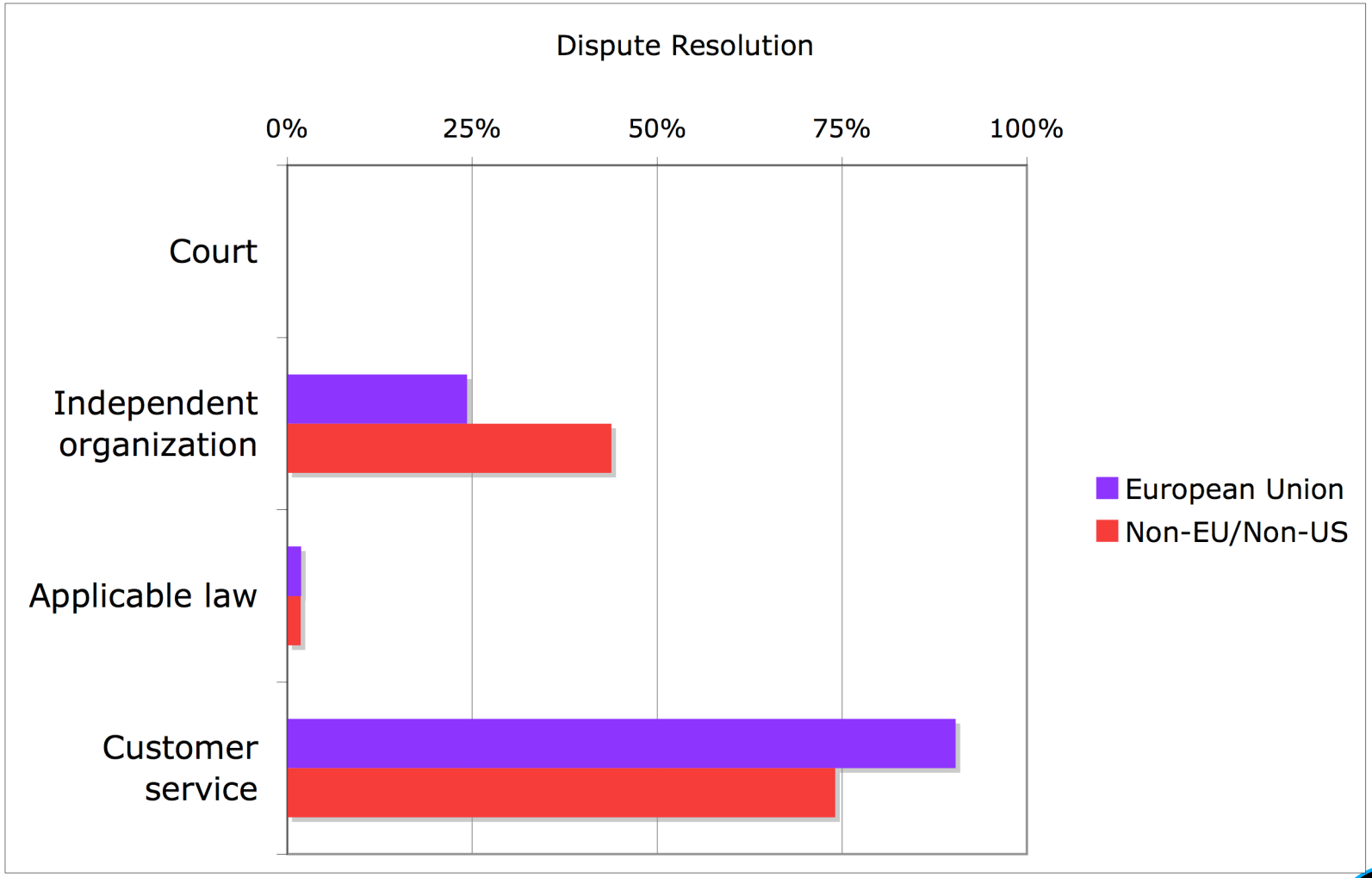
Data recipients (EU v. Non-EU/Non-US)



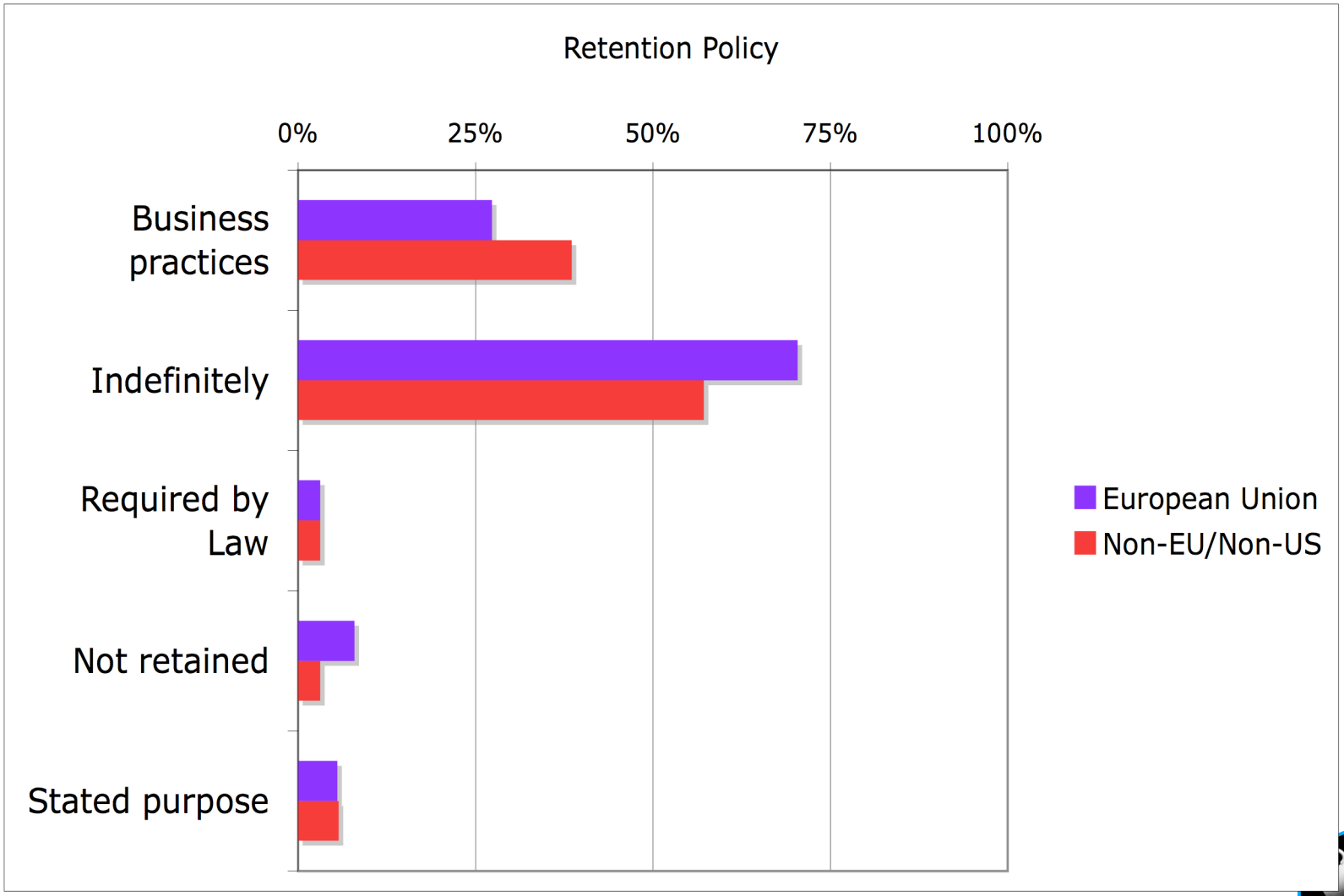
Access provisions (EU v. Non-EU/Non-US)



Dispute resolution (EU v. Non-EU/Non-US)

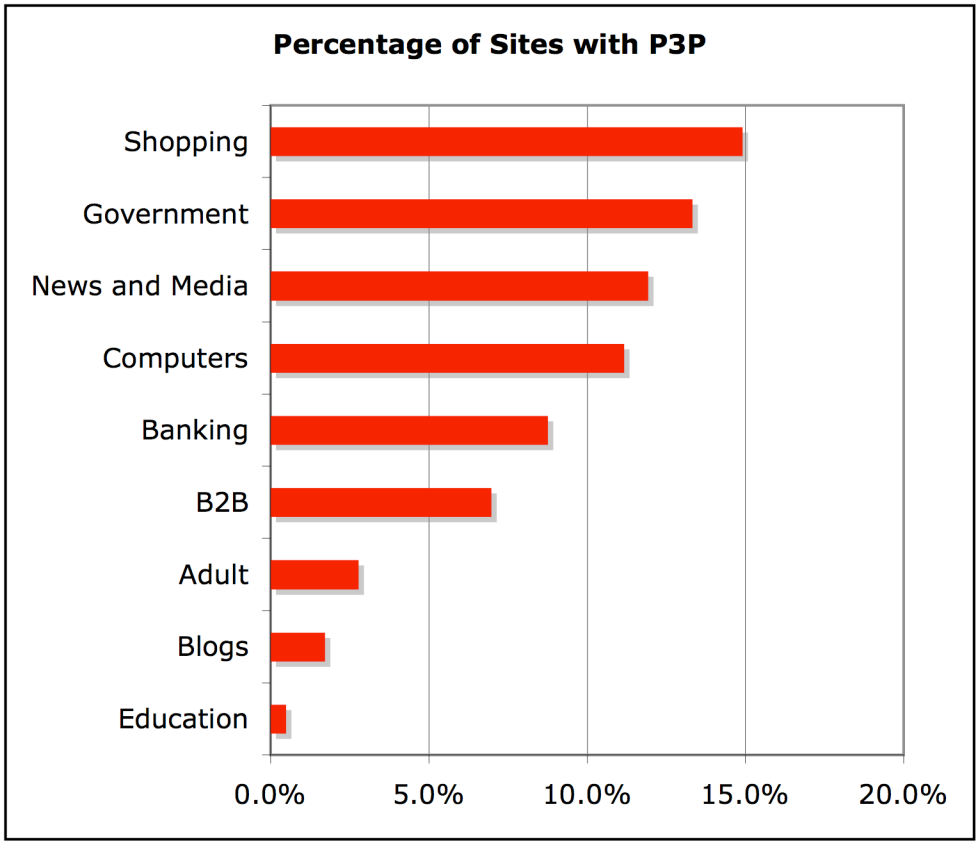


Data retention policy (EU v. Non-EU/Non-US)



P3P-enabled sites by category

- 16,919 sites categorized using Yahoo! categories



Rate of change of P3P policies

- Over an eight week span (10/25/06 - 12/20/06) we observed the following changes:
  - 69 policy changes
  - 70 policy "removals"
    - Sites temporarily or permanently unavailable
    - Sites that put up robots files that do not permit retrieval of P3P policy
    - (no sites appear to have explicitly removed P3P policy)
  - 470 new policies, of which 272 are unique
- Net growth rate of 4.16% annually

P3P Errors

- Syntactic errors
  - Policies do not follow the P3P specification
  - Some errors are "critical"
    - These policies cannot be parsed
  - Other errors are non-critical
    - We can determine the intent of these policies
- Semantic errors
  - P3P policy and natural language policy disagree

P3P Syntactic Errors

Error	Popular	Privacy Finder
Old Version	15 (71.4%)	9,155 (62.2%)
No Policy Name	3 (14.3%)	6,289 (42.7%)
No Errors	6 (28.6%)	4,014 (27.3%)
Policy Validation Error	1 (4.8%)	1,157 (7.9%)
Bad XML Root	8 (38.1%)	1,125 (7.6%)
Policy Expired	0	474 (3.2%)
Policy Vocabulary Error	0	453 (3.1%)
No Policy Elements	0	252 (1.7%)
Incorrect XML	0	204 (1.4%)
Policy Access Error	1 (4.8%)	183 (1.2%)
No Namespace	0	151 (1.0%)
Malformed INCLUDE/EXCLUDE	0	56 (0.4%)
No <META/> Tag	0	21 (0.1%)
No Policy Found	0	5 (0%)
Not A Policy	0	2 (0%)
Total Policies	**21**	**14,720**

P3P Syntactic Errors

- Most (over 71%) policies have errors
  - Many are non-critical
- Most (~70%) errors are due to old versions
  - Little change in syntax
  - Many policies missing the name fragment

Semantic Errors

	<ACCESS> (1)	<CATEGORIES> (17)	<DISPUTES> (4)	<NON-IDENTIFIABLE> (1)	<PURPOSE> (12)	<RECIPIENT> (5)	<REMEDIES> (1)	<RETENTION> (1)	Total
1. yahoo.com	0	3	0	0	2	4	0	0	9
2. geocities.com	0	3	0	0	2	4	0	0	9
3. hotmail.com	1	1	0	0	2	0	0	1	5
4. superpages.com	1	6	0	0	5	3	0	1	16
5. angelfire.com	0	0	0	0	3	1	0	0	4
6. walmart.com	0	4	1	0	4	2	1	1	13
7. go.com	0	1	0	0	3	2	0	1	7
8. microsoft.com	0	2	0	0	2	0	0	0	4
9. ticketmaster.com	1	7	0	1	5	3	0	0	17
10. usps.com	0	1	0	0	3	2	1	1	8
11. dealtime.com	1	7	1	0	5	1	1	1	17
12. rootsweb.com	1	5	0	0	5	2	0	1	14
13. hgtv.com	1	3	0	0	1	3	0	1	9
14. wachovia.com	0	5	0	0	5	1	1	1	13
15. tripod.com	0	0	0	0	3	1	0	0	4
16. sportsline.com	0	6	0	0	2	3	0	1	12
17. qvc.com	1	7	0	0	4	1	0	0	13
18. download.com	0	5	1	0	2	5	0	0	13
19. usatoday.com	0	2	1	0	1	1	1	0	6
20. about.com	1	4	2	0	4	2	0	1	14
21. wunderground.com	1	7	0	0	3	0	0	0	11
Policies with Error	**9**	**19**	**5**	**1**	**21**	**18**	**5**	**11**	**217**

- We compared P3P policies with our manual coding
- Differences may be due to one policy being more specific
- Some differences are conflicts indicating an error in one or both policies

Semantic Errors

- <ACCESS> errors
  - 43% specified different policies for reviewing personal information
- <CATEGORIES> errors
  - 80% showed different information collected
- <DISPUTES> errors
  - Very few
  - Most due to non-reporting of third parties (e.g. TrustE)
- <NON-IDENTIFIABLE> errors
  - Only ticketmaster.com used this incorrectly

Semantic Errors

- <PURPOSE> errors
  - 8 policies mention marketing not mentioned in P3P policy
  - 5 policies mention telemarketing not mentioned in natural language policy
- <RECIPIENT> errors
  - Significant differences
  - 15 natural language policies specify recipients not in the P3P policies
  - 3 P3P policies overly report sharing
  - 3 P3P policies accurately report sharing

Semantic Errors

- <RETENTION> errors
  - No natural language policies mention retention
    - Eleven sites require it from their P3P policies

Error Summary

- Syntactic errors show P3P policies not being adequately checked and maintained
- Semantic errors show misunderstanding of P3P
  - Wachovia claims not to use P3P
    - P3P policy is available
    - Only references cookies from their site
  - <financial> tag overused
    - Only meant for information beyond purchases
  - Many overly-broad P3P policies
    - Five policies mention concerns not in their natural language policies
      - Worst case scenarios
      - Benefits users
- Errors have little affect on high/medium/low settings in Privacy Finder and Privacy Bird

We want your feedback

- What parts of this report are most useful to you?
- What would you like to see us update next year?
- What new data would you like to see us collect next year?
- Collecting P3P data is easy, coding natural language policies is hard - does data on natural language policies provide value to you?

CMU Usable Privacy and Security Laboratory

http://cups.cs.cmu.edu/

Carnegie Mellon