

Examining the Content and Privacy of Web Browsing Incidental Information

Kirstie Hawkey
Kori Inkpen

EDGE | LAB

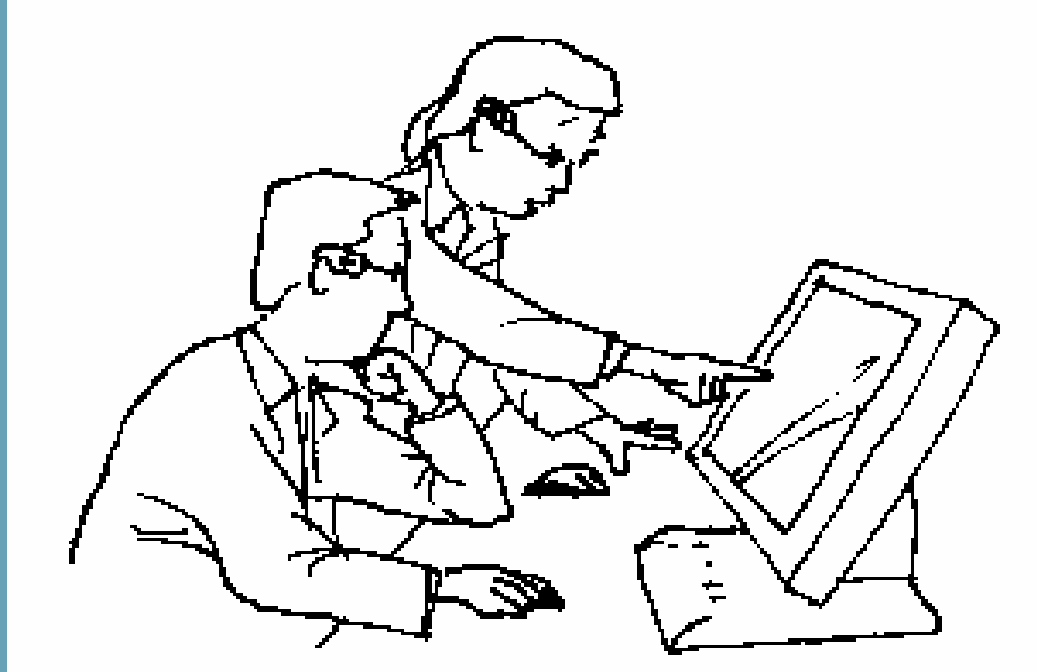


**DALHOUSIE
UNIVERSITY**

Inspiring Minds



Incidental Information Privacy



Traces of previous activity visible on personal computer display

Privacy issues arise when others can view your display.

The information, incidental to the task at hand, may not be appropriate for current viewing context



Navigation toolbar with icons for Back, Forward, Stop, Refresh, Home, Search, Favorites, RSS, Mail, Print, Send, Attachments, Phone, Bookmarks, and People.

Address <http://www.google.ca/> Go

Search bar with Google logo, Search Web dropdown, PageRank, 331 blocked, AutoFill, and Options.

- History ×
- View ⌵ Search
- odetocode (odetocode.c...)
 - office.microsoft (office.mi...)
 - pandab (www.pandab.org)
 - pcmag (www.pcmag.com)
 - pebbles.hcii.cmu (www.p...)
 - physicsforums (www.phy...)
 - portal.acm (portal.acm.org)
 - python (www.python.org)
 - qualityforge (qualityforge...)
 - radiks (www.radiks.net)
 - realityblurred (www.realit...)
 - realitytvworld (www.reali...)
 - rustemsoft (rustemsoft.c...)
 - sanitycheck (www.sanity...)
 - scwm.sourceforge (scwm...)
 - sea.search.msn (sea.sea...)
 - search.discovery (search...)
 - search.microsoft (search...)



Web [Images](#) [Groups](#) [News](#) [Local](#)^{New!} [Desktop](#) [more »](#)

Search input: p|

- panic attack during public speaking
- personal bankruptcy laws
- persperation + nervous
- perspiration + nervous
- presentation anxiety
- privacy research**
- private web browsing

[Advanced Search](#)
[Preferences](#)
[Language Tools](#)

[Advertising Programs](#) - [Business Solutions](#) - [About Google](#) - [Go to Google.com](#)

©2005 Google - Searching 8,058,044,651 web pages

Windows taskbar with Start button, application icons (Internet Explorer, Outlook, etc.), and system tray showing time 10:30 AM.



Privacy Management

Systems approach

1. Classify content as created with privacy level
2. Filter content appropriately according to viewing context

Our previous work indicates manual classification by users would be difficult

- Large number of sites, rapid bursts of browsing

An automated approach may be to use the content category of the web page

- Commercial content filtering products (e.g. Cerberian)



Research Questions

1. How does the content of visited web pages affect participants' privacy classifications?
2. Is an automated approach to content classification scheme feasible?



Participants

- Recruited from Dalhousie University community
 - 11 students / 4 office staff
 - 10 female / 5 male
 - Average age 27.8 (18 to 44)
- Mixture of technical and non-technical, desktop and laptop users
- Reported usual reasons for web browsing
 - 37% personal browsing
 - 18% work-related
 - 45% school-related



Methodology

- Week long field study
- Browser Helper Object
- Logged data included:
 - Browser window ID
 - Date/Time stamp
 - Page title, URL

Electronic Diary

- 4-level privacy scheme
- Selectively sanitized data

Select rows, then click privacy level

Public Semi-Public Private Don't Save

Window ID	Date / Time	Page Title	URL	Privacy Level
263880	3/15/2005 00:08:43:25	Google	http://www.google.ca/	Don't Save
264016	3/15/2005 00:08:43:09	Google	http://www.google.ca/	Don't Save
264016	3/15/2005 00:09:07:69	Google Search: c	http://www.google.com/search?sourceid=navclien	Public
264454	3/15/2005 00:08:42:15	Google	http://www.google.ca/	Don't Save
329560	3/15/2005 00:08:38:90	Google	http://www.google.ca/	Don't Save
229560	3/15/2005 00:09:02:82	Google Search: b	http://www.google.com/search?sourceid=navclien	Public
461094	3/15/2005 00:08:42:83	Google	http://www.google.ca/	Don't Save
461094	3/15/2005 00:09:12:79	Google Search: d	http://www.google.com/search?sourceid=navclien	Public
9765648	3/15/2005 00:08:41:79	Google	http://www.google.ca/	Don't Save
3408780	3/15/2005 00:14:17:67	Google	http://www.google.ca/	Don't Save
3408780	3/15/2005 00:14:24:34	Google Search: d	http://www.google.com/search?sourceid=navclien	Public
3539624	3/15/2005 00:24:13:74	Google	http://www.google.ca/	Don't Save
3539624	3/15/2005 00:24:19:35	Google Search: stacey scott	http://www.google.com/search?sourceid=navclien	Semi-Public
3539624	3/15/2005 00:24:31:86	Google Search: stacey scott denfence	http://www.google.com/search?sourceid=navclien	Semi-Public
3539624	3/15/2005 00:24:38:47	Google Search: stacey scott defence	http://www.google.com/search?hl=en&rls=GGLD	Semi-Public
3539624	3/15/2005 00:24:51:58	Google Search: stacey scott defence calgary	http://www.google.com/search?hl=en&rls=G	Semi-Public
132134	3/15/2005 08:38:45:90	Google	http://www.google.ca/	(null)
132134	3/15/2005 08:39:15:95	zz-Sanitized-zz	zz-search for medical info-zz	Private
132134	3/15/2005 08:40:03:08	zz-Sanitized-zz	zz-search for medical info-zz	Private
132134	3/15/2005 08:40:25:33	zz-Sanitized-zz	zz-search for medical info-zz	Private
132134	3/15/2005 08:40:39:02	zz-Sanitized-zz	zz-search for medical info-zz	Private
132134	3/15/2005 08:40:56:95	zz-Sanitized-zz	zz-search for medical info-zz	Private
132134	3/15/2005 08:41:14:76	zz-Sanitized-zz	zz-search for medical info-zz	Private
132134	3/15/2005 08:44:33:16	zz-Sanitized-zz	zz-search for medical info-zz	Private
197892	3/15/2005 09:27:52:95	Google	http://www.google.ca/	(null)
197892	3/15/2005 09:28:00:92	Canada411	http://www.canada411.com	(null)
197892	3/15/2005 09:28:03:39	http://canada411.yellowpages.ca/	http://canada411.yellowpages.ca/	(null)
197892	3/15/2005 09:28:03:68	Canada411	http://canada411.yellowpages.ca/searchBusiness.	(null)
197892	3/15/2005 09:28:20:39	Canada411	http://canada411.yellowpages.ca/searchBusiness.	(null)
197892	3/15/2005 09:28:22:43	Canada411	http://canada411.yellowpages.ca/searchBusiness.	(null)
1705634	3/15/2005 11:56:27:34	Google	http://www.google.ca/	(null)
1705634	3/15/2005 11:58:12:56	http://www.google.ca/search?hl=en&q=backup+r	http://www.google.ca/search?hl=en&q=backup+r	(null)
1705634	3/15/2005 11:58:22:63	Backing up the Windows registry	http://service1.symantec.com/SUPPORT/tsgeninfo	(null)
1705634	3/15/2005 11:59:10:99	Google Search: backup registry	http://www.google.ca/search?hl=en&q=backup+r	(null)
1705634	3/15/2005 11:59:24:80	Google Search: windows registry copy	http://www.google.ca/search?hl=en&c2coff=1&q	(null)
1705634	3/15/2005 11:59:38:07	Windows Registry help	http://www.computerhope.com/registry.htm	(null)

Hide URL info

Sanitize

Before you exit the diary.

Create Privacy Gradient Report

Content Categories

- 55 commercial web filtering categories (Cerberian)
- Theoretical privacy classification task

Give a classification for each of these types of websites based on whether or not you would mind if others saw that you visited a site of this type (either accidentally or on purpose). Classify it as “public”, if you wouldn’t mind anybody seeing it, “semi-public” if you wouldn’t mind some subset of people seeing it, “private” if you would like to restrict most others from seeing it but still want to have access to it yourself, and “don’t save” if you would not want a site of this type saved by your web browser.

Category Name / Examples	Description	Classification
Adult/Mature Content www.humorbomb.org www.steakandcheese.com	Sites that contain material of adult nature that does not necessarily contain excessive violence, sexual content, or nudity. These sites include very profane or vulgar content and sites that are not appropriate for children.	<input type="checkbox"/> Public <input type="checkbox"/> Semi-public <input type="checkbox"/> Private <input type="checkbox"/> Don't Save
Pornography www.playboy.com www.whitehouse.com	Sites that contain sexually explicit material for the purpose of arousing a sexual or prurient interest.	<input type="checkbox"/> Public <input type="checkbox"/> Semi-public <input type="checkbox"/> Private <input type="checkbox"/> Don't Save
Sex Education www.viagra.com www.sexuality.org	Sites that provide graphic information (sometimes graphic) on reproduction, safe sex practices, sexuality, birth control, and sexual development. Also includes sites that offer tips for better sex as well as products used for sexual enhancement.	<input type="checkbox"/> Public <input type="checkbox"/> Semi-public <input type="checkbox"/> Private <input type="checkbox"/> Don't Save
Intimate Apparel/Swimsuit www.victoriassecret.com www.fredericks.com	Sites that contain images or offer the sale of swimsuits or intimate apparel or other types of suggestive clothing. Does not include sites selling undergarments as a subsection of other products offered.	<input type="checkbox"/> Public <input type="checkbox"/> Semi-public <input type="checkbox"/> Private <input type="checkbox"/> Don't Save
Nudity	Sites containing nude or seminude depictions of the human body. These depictions are not necessarily sexual in intent or effect, but may include sites	<input type="checkbox"/> Public <input type="checkbox"/> Semi-public



Content Category Analysis

- Researchers partitioned participants' actual browsing from the week into categories
 - Same 55 Cerberian categories
 - Combined all participant data (31,160 page visits)
 - Sorted by URL
 - Filtered URLs with Zone Alarm Security Suite's parental control feature
 - Manual classification of remainder



Results

Visited Categories Varied

41/55 categories (average 21, 15 to 29)

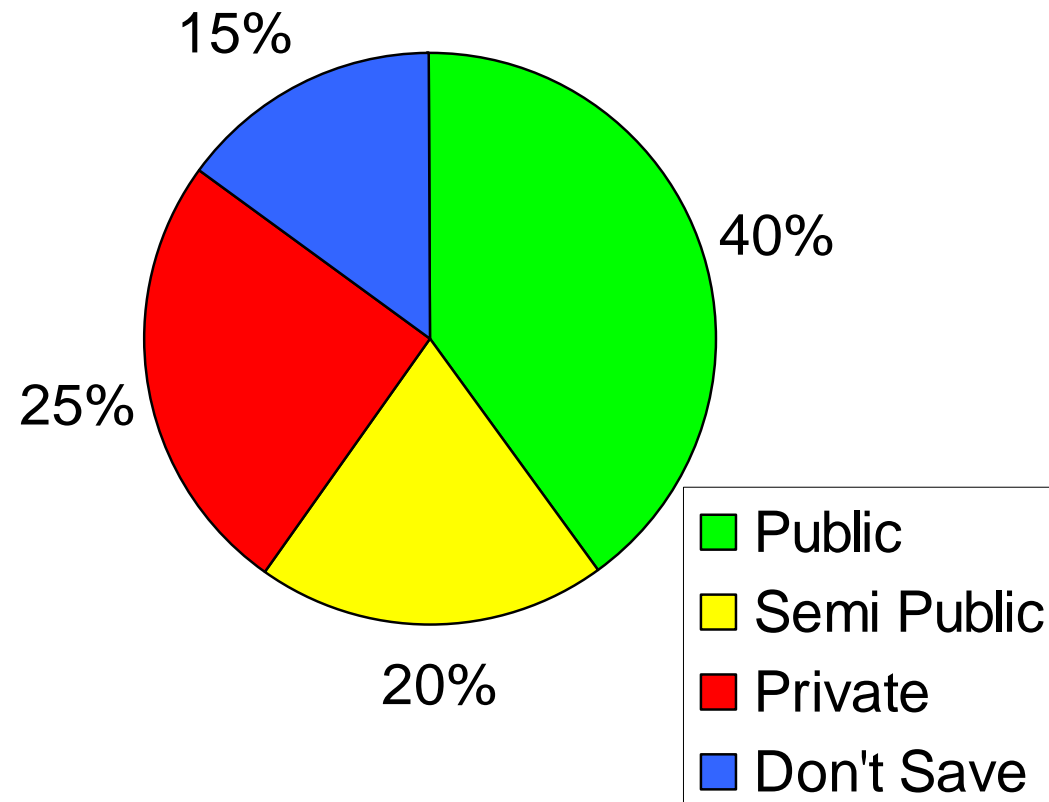
Category	# pages	# participants	# part. with 10+ pages
Search Engines/Portals	6310	15	15
Education	3315	15	14
Email	5082	14	14
Reference	2055	14	13
News/Media	1320	14	7
Shopping	770	14	10
Arts/Entertainment	665	14	12
Society/Lifestyle	1136	13	8
...			
News Group	1303	9	3

Privacy Levels Applied (Overall)

How do privacy levels change according to category of content?

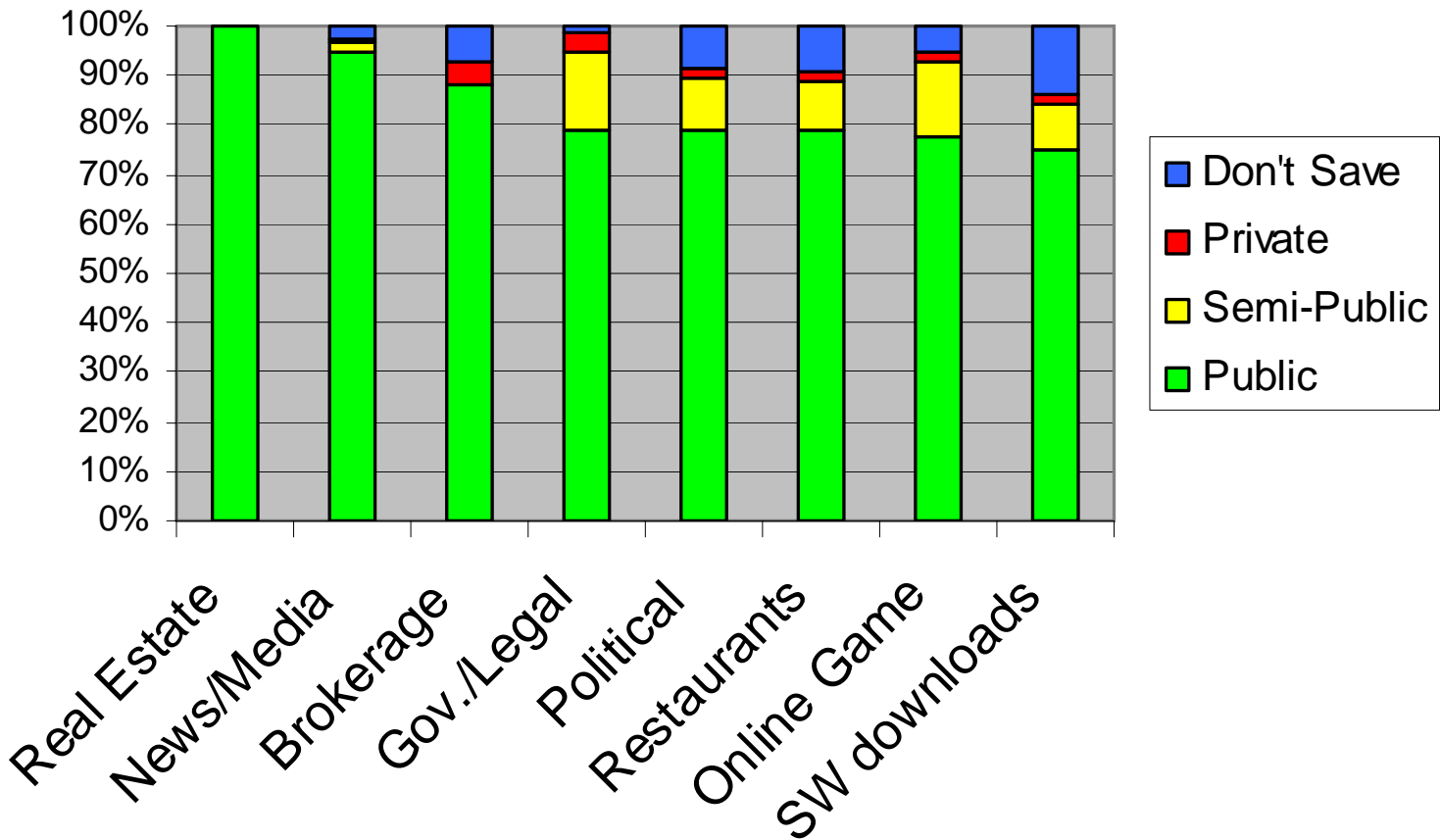
■ K-means cluster analysis found 5 clusters

- public
- semi-public
- private
- public/don't save
- mixture



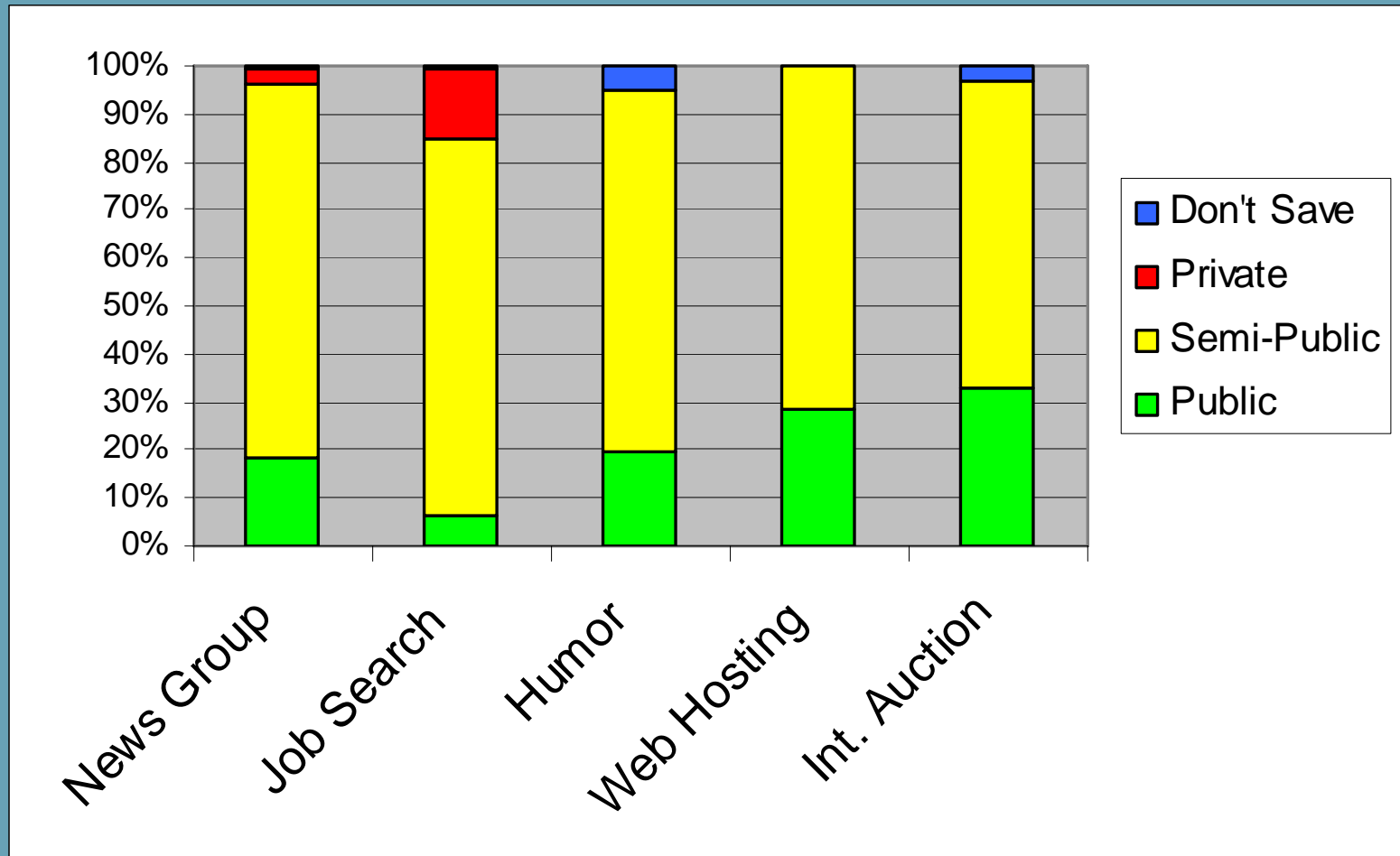
Cluster: public

9.8% of visited pages



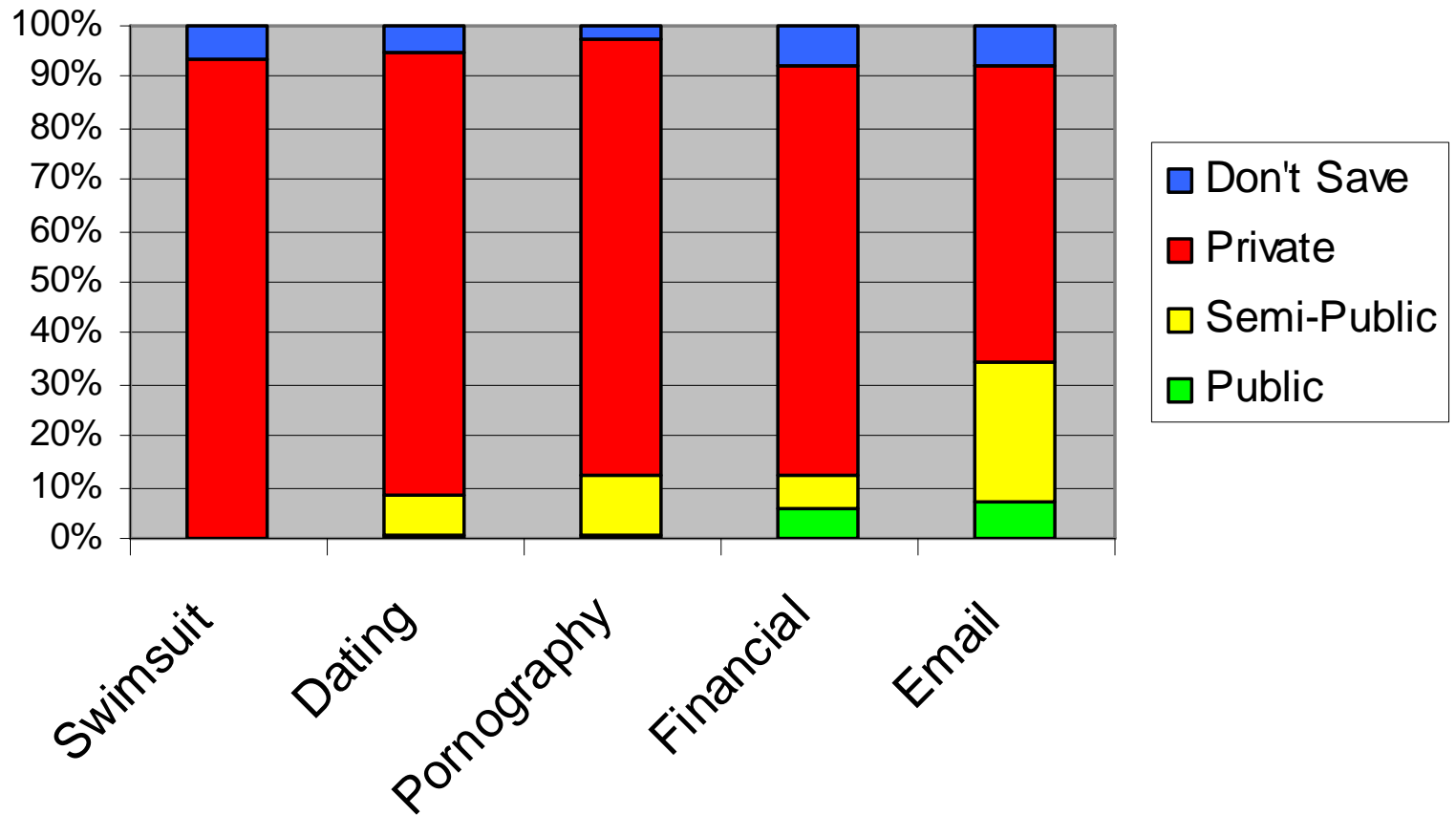
Cluster: semi-public

6.4% of visited pages



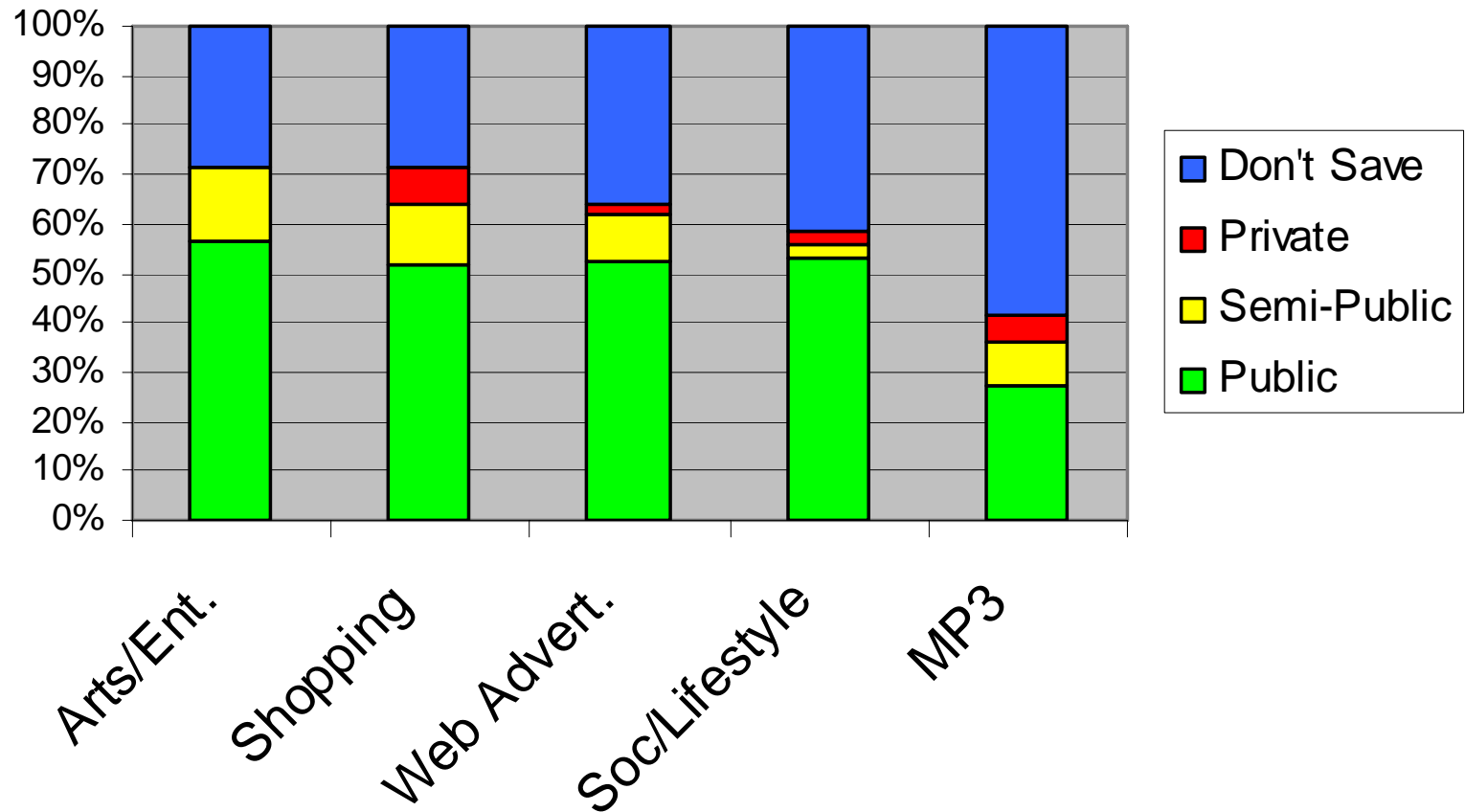
Cluster: private

21.0% of visited pages



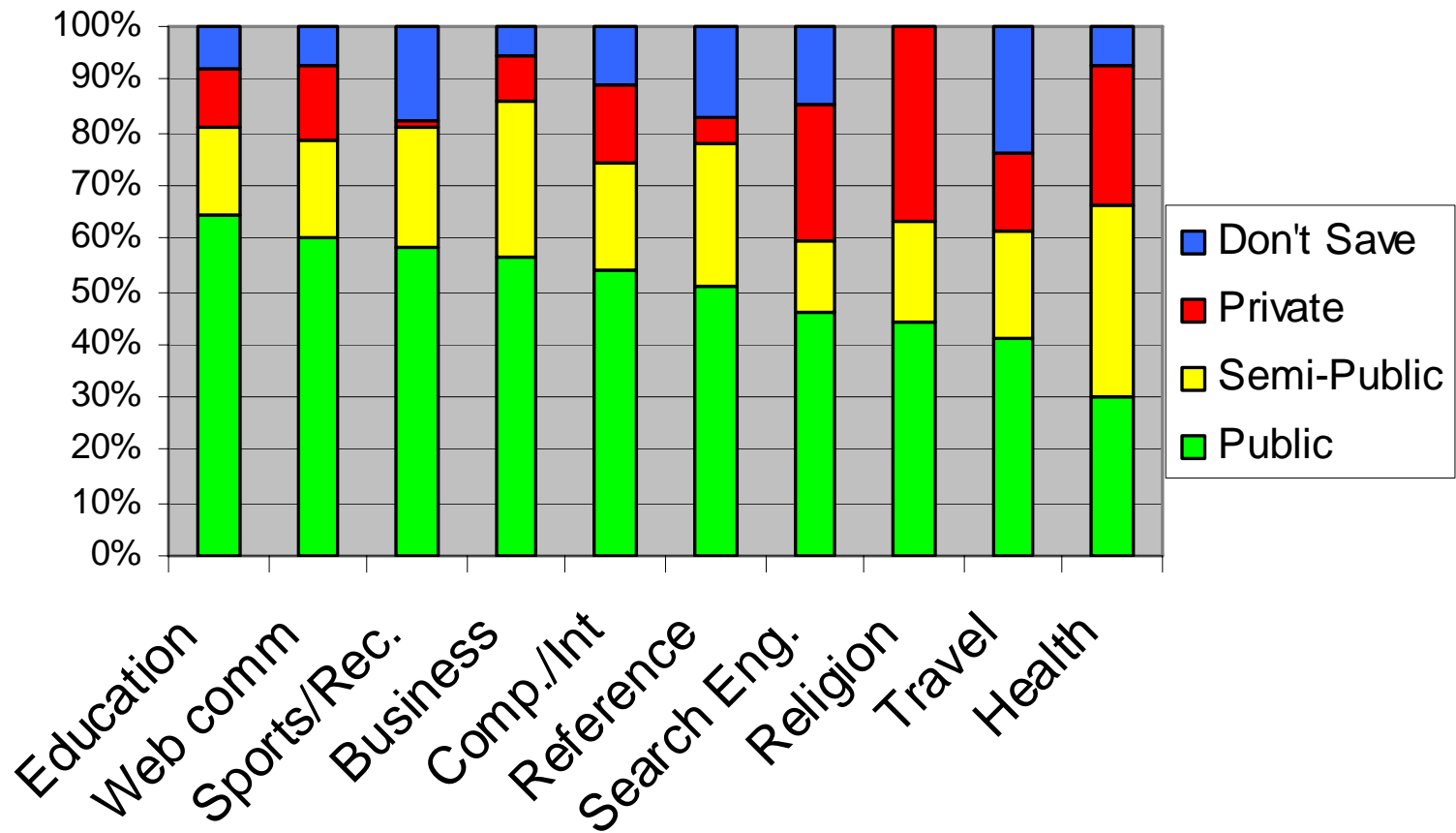
Cluster: public/don't save

9.2% of visited pages



Cluster: Mixture

44.1% of visited pages





Possible Classification Approaches

- **Standardized approach**
 - Common default privacy level for categories
 - General consensus needed as to which privacy level is appropriate for each content category
- **Personalized approach**
 - User defined default privacy level for categories
 - Individuals need to be fairly consistent at their desired privacy level within each category
 - Individuals must be able to specify default privacy levels for each category



Evaluate Standardized Approach

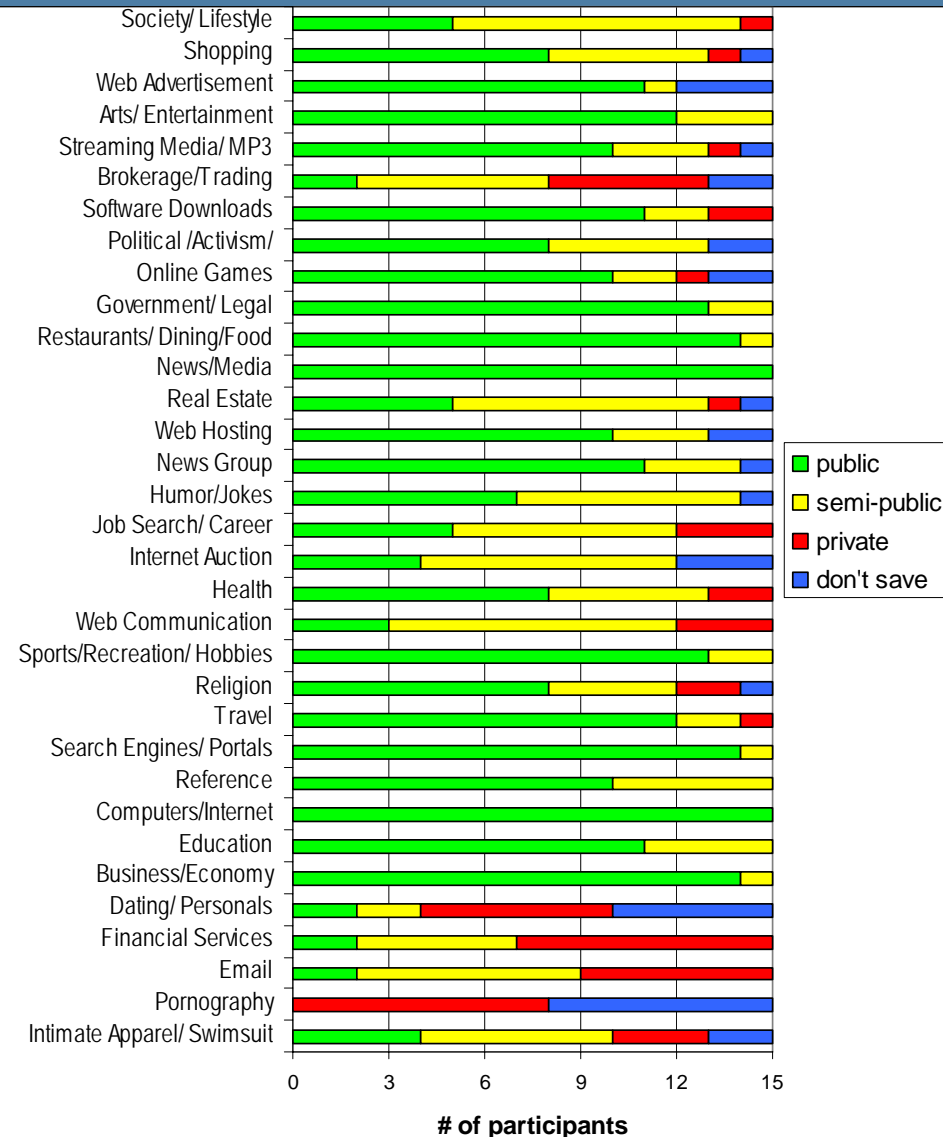
Examine consistency between participants in their theoretical content category classification task

Examine consistency between participants in their privacy classification of visited pages within a category

Theoretical Classification Task

Little agreement about appropriate privacy level

- Only 8 categories with 80% (12+ participants) agreement
- Only 2 categories in complete agreement





Actual Privacy Classifications

How much agreement is there between participants within each category?

- 30 categories had 2+ participants with 10+ page visits
- Determined primary privacy level for each participant for each category

Only 4/30 categories had complete agreement between participants

- News/media, political activism, pornography, web hosting



Feasibility: Standardized Approach

Is a standardized approach to automated privacy classification based on content category feasible?

- No

- Clustering showed basic agreement for some categories (C2: Public, C3: Semi-Public, C5: Private), but C2: Public/Don't Save and C4: Mixture accounted for 53.3% of visited pages
- Low consistency between participants in primary privacy level applied
- Theoretical web category classification task showed little agreement for appropriate classifications



Evaluate Personalized Approach

Examine participant consistency at applying a single privacy level to page visits within a category

Examine ability of participants to predict which privacy level they will apply

Consistency Within a Category

How consistent were participants in assigning privacy levels to pages within a category (regardless of their primary privacy level)?

- For each participant with 10+ page visits in a category we computed a normalized consistency:

$$\text{Norm. consistency} = \frac{\# \text{ pages at primary privacy level}}{\text{total page visits in category}}$$

- Category consistency is average of participant consistency

Consistency Within a Category

■ Average 81% consistency

Most Consistent Categories	%	Least Consistent Categories	%
Real Estate	100	Search engines/portals	61
Restaurants/dining/food	99	Education	65
Intimate apparel/swimsuits	97	Computers/internet	66
News/media	96	Web advertisements	71
Political/activism/advocacy	95	Reference	76
Brokerage/trading	95	Web communication	76
Society/lifestyle	93	Streaming media/mp3	76
Health	92	News group	78
Internet Auction	92	Religion	78
Sports/recreation/Hobbies	91	Humour/Jokes	79



Prediction Accuracy

How well did participants predict what privacy levels they would apply to a category of web browsing?

- Compared participants theoretical content classification with privacy levels they applied to their web browsing
- For each category, we computed participants' accuracy:

$$\text{Accuracy} = \frac{\# \text{ pages at predicted privacy level}}{\text{total page visits in category}}$$

Prediction Accuracy

■ Average 58% accuracy

Most Accurate Categories	%	Least Accurate Categories	%
Real Estate	99	Brokerage/trading	0
Intimate apparel/swimsuits	95	Society/lifestyle	10
News/media	95	Health	16
Internet auction	95	Dating/personals	18
Restaurants/dining/food	88	Web hosting	29
Job search/career	86	Web communication	32
Pornography	86	Shopping	38
Government/legal	78	Sports/recreation/hobbies	39
Email	77	Religion	44
Financial Services	75	Travel	45



Feasibility: Personalized Approach

Is a personal privacy management system using automated privacy classification based on content category feasible?

- Maybe

- Participants were consistent within many categories

- 12/34 had greater than 90% consistency

- BUT 13/34 had less than 80% consistency

- Prediction accuracy varied greatly both across participants and for different content categories



Reasons for Inconsistencies


- Dual nature of Don't Save
- Semi-public ("it depends")
 - Uncertainty about appropriate classification may be due to potential viewers and also page content
 - Viewing context may be partially resolved when considering actual page content



Reasons for Inconsistencies

Category characteristics

- General categories
 - Specific pages can have very different content
- Varying task purposes
 - Information or transaction?
 - Login, https
- Complex/dynamic pages
 - Privacy sensitivity may vary depending on content at a given time



Recommendations to Improve Accuracy

Refine content categorization through heuristics

- Keywords
- Login / secure site
- Query string

More effectively communicate category characteristics to users

- Include examples of the types of content and activities that may be visible



Summary

A standardized approach is not feasible

- Inconsistencies between participants

Personalized scheme may be feasible

- participants were fairly consistent within most categories

BUT

- More fine-grained approach to content classification is required
- Users would need richer descriptions of categories

Thanks to:

- NSERC
- NECTAR
- Dalhousie University
- EDGE Lab



Kirstie Hawkey
hawkey@cs.dal.ca

EDGE | LAB



**DALHOUSIE
UNIVERSITY**

Inspiring Minds

