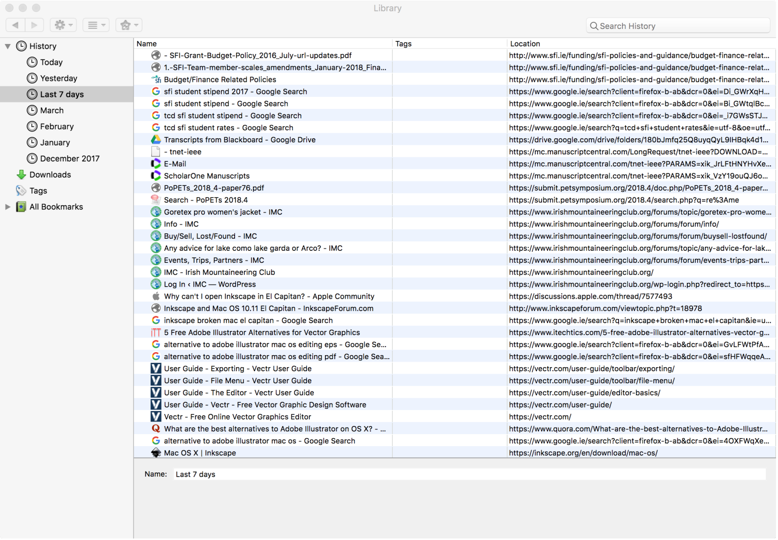
Learning & Privacy in Online Search

Doug Leith

School of Computer Science & Statistics,
Trinity College Dublin

Our Web Search History ...

... Its full of little details of our lives, work, interests, plans



Library

Search History

History
Today
Yesterday
Last 7 days
March
February
January
December 2017
Downloads
Tags
All Bookmarks

Name	Tags	Location
- SFI-Grant-Budget-Policy_2016_July-url-updates.pdf		http://www.sfi.ie/funding/sfi-policies-and-guidance/budget-finance-relat...
1.-SFI-Team-member-scales_amendments_January-2018_Fina...		http://www.sfi.ie/funding/sfi-policies-and-guidance/budget-finance-relat...
Budget/Finance Related Policies		http://www.sfi.ie/funding/sfi-policies-and-guidance/budget-finance-relat...
sfi student stipend 2017 - Google Search		https://www.google.ie/search?client=firefox-b-ab&dcr=0&ei=Di_GWrXqH...
sfi student stipend - Google Search		https://www.google.ie/search?client=firefox-b-ab&dcr=0&ei=Bi_GWtqIBc...
tcd sfi student stipend - Google Search		https://www.google.ie/search?client=firefox-b-ab&dcr=0&ei=_i7GWsSTJ...
tcd sfi student rates - Google Search		https://www.google.ie/search?q=tcd+sfi+student+rates&ie=utf-8&oe=utf...
Transcripts from Blackboard - Google Drive		https://drive.google.com/drive/folders/180bJmfq25Q8uyqQyL9IHBqk4d1...
- tnet-ieee		https://mc.manuscriptcentral.com/LongRequest/tnet-ieee?DOWNLOAD=...
E-Mail		https://mc.manuscriptcentral.com/tnet-ieee?PARAMS=xik_JrLFtHNYHvXe...
ScholarOne Manuscripts		https://mc.manuscriptcentral.com/tnet-ieee?PARAMS=xik_VzY19ouQJ6o...
PoPETs_2018_4-paper76.pdf		https://submit.petsymposium.org/2018.4/doc.php/PoPETs_2018_4-paper...
Search - PoPETs 2018.4		https://submit.petsymposium.org/2018.4/search.php?q=re%3Ame
Goretex pro women's jacket - IMC		https://www.irishmountaineeringclub.org/forums/topic/goretex-pro-wome...
Info - IMC		https://www.irishmountaineeringclub.org/forums/forum/info/
Buy/Sell, Lost/Found - IMC		https://www.irishmountaineeringclub.org/forums/forum/buysell-lostfound/
Any advice for lake como lake garda or Arco? - IMC		https://www.irishmountaineeringclub.org/forums/topic/any-advice-for-lak...
Events, Trips, Partners - IMC		https://www.irishmountaineeringclub.org/forums/forum/events-trips-part...
IMC - Irish Mountaineering Club		https://www.irishmountaineeringclub.org/
Log In ‹ IMC — WordPress		https://www.irishmountaineeringclub.org/wp-login.php?redirect_to=https...
Why can't I open Inkscape in El Capitan? - Apple Community		https://discussions.apple.com/thread/7577493
Inkscape and Mac OS 10.11 El Capitan - InkscapeForum.com		http://www.inkscapeforum.com/viewtopic.php?t=18978
inkscape broken mac el capitan - Google Search		https://www.google.ie/search?q=inkscape+broken+mac+el+capitan&ie=u...
5 Free Adobe Illustrator Alternatives for Vector Graphics		https://www.itechtics.com/5-free-adobe-illustrator-alternatives-vector-g...
alternative to adobe illustrator mac os editing eps - Google Se...		https://www.google.ie/search?client=firefox-b-ab&dcr=0&ei=GvLFWtPfA...
alternative to adobe illustrator mac os editing pdf - Google Sea...		https://www.google.ie/search?client=firefox-b-ab&dcr=0&ei=sfHFWqqeA...
User Guide - Exporting - Vectr User Guide		https://vectr.com/user-guide/toolbar/exporting/
User Guide - File Menu - Vectr User Guide		https://vectr.com/user-guide/toolbar/file-menu/
User Guide - The Editor - Vectr User Guide		https://vectr.com/user-guide/editor-basics/
User Guide - Vectr - Free Vector Graphic Design Software		https://vectr.com/user-guide/
Vectr - Free Online Vector Graphics Editor		https://vectr.com/
What are the best alternatives to Adobe Illustrator on OS X? - ...		https://www.quora.com/What-are-the-best-alternatives-to-Adobe-Illustr...
alternative to adobe illustrator mac os - Google Search		https://www.google.ie/search?client=firefox-b-ab&dcr=0&ei=4OXFWqXe...
Mac OS X | Inkscape		https://inkscape.org/en/download/mac-os/

Name: Last 7 days

What About Privacy ? And Who Cares Anyway ?

- "I've nothing to hide"
- "More personalised search and ads helps me" ?

What About Privacy ? And Who Cares Anyway ?

- “I’ve nothing to hide”
- “More personalised search and ads helps me” ?
- Ignorance is bliss – unconcerned because unaware ?
- What if its your teacher who is looking at it ? Or your parents ?
- Analogy with smoking ?

Its Not Just About You

- Impact of privacy breaches on business reputation e.g.
  - Facebook share value[1] (Cambridge Analytica etc)
  - Google's usage of UK health data[2]
- Business to business leakage of information e.g.
  - Pre-patent due diligence searches
- Abuses e.g.
  - Getting the dirt on the political opposition
  - Chilling effect on debate, self-censorship (try talking to someone who has lived under an oppressive regime)
  - Wilful misunderstandings

---

[1] E.g see "Facebook Suffers Worst-Ever Drop in Market Value", Wall Street Journal, July 26 2018

[2] E.g. see "Google given access to healthcare data of up to 1.6 million patients", Guardian 4 May 2016

Privacy

Privacy vs Security

- Privacy is not the same as security/crypto
  - Security $\rightarrow$ admission control, akin to a lock. Either have access or do not
  - Privacy $\rightarrow$ release information in return for a benefit. Expect a trade-off between amount of information released and amount of benefit received.
- Current approach to online information sharing is largely binary, however – all in or all out.
  - Of course, this is by design and intentional.
  - But is it really necessary for delivery of personalised services (separately from the third party market in our data) ?

Privacy

Importance of Verification

- Contracts cannot be enforced unless compliance (or otherwise) can be verified.
- Suppose a provider says it doesn't use data that it sees, or that it provides some sort of private service. How can we verify this claim ? If we can't then we can't detect non-compliance let alone enforce compliance.
- Profit motive may be helpful here, hopefully at least.
- Note that we can only hope to provide evidence of learning, cannot prove absence of learning.

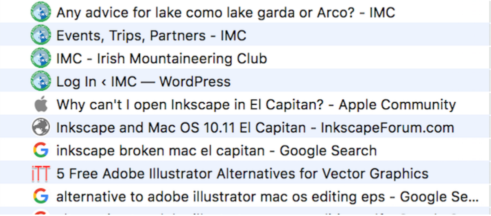
Online Search

Interface: Stuctured Interaction

Structured interface:

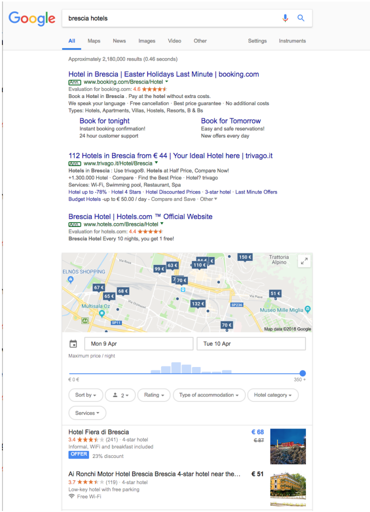
- Input queries
- Service responds with a results page
- User can click on links

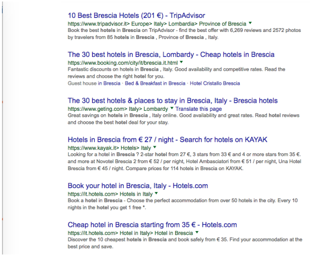
We tend to enter sequences of related queries on a particular topic, then change topic and repeat. So interactions tend to consist of a set of sessions, each on a particular topic.



Online Search

Interface: Response Page





Online Search

Interface: Response Page

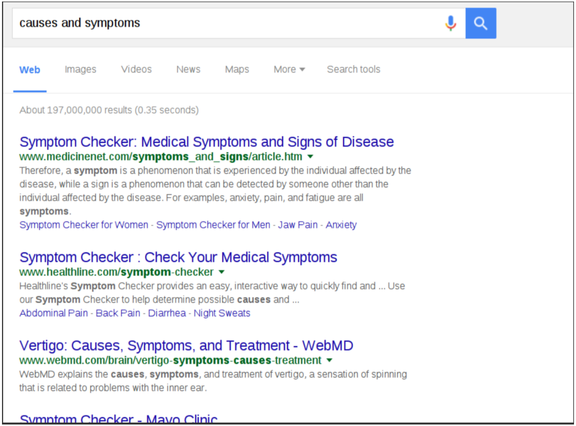
- Search results tend to be invariant ...
- ... but adverts are changeable.

	Bing		Google	
Topic	**Advert**	**Link**	**Advert**	**Link**
anorexia	65.4% ± 7.7%	3.6% ± 0.3%	34.8% ± 1.5%	0.9% ± 0.2%
bankrupt	15.8% ± 1.5%	5.0% ± 0.3%	39.0% ± 2.5%	2.0% ± 0.3%
diabetes	49.4% ± 12.5%	3.9% ± 0.3%	39.5% ± 1.7%	0.9% ± 0.2%
disabled	12.4% ± 1.0%	3.5% ± 0.2%	17.3% ± 1.7%	2.1% ± 0.3%
divorce	15.8% ± 1.7%	4.7% ± 0.4%	22.1% ± 2.5%	2.9% ± 0.5%
gambling	15.7% ± 1.3%	4.0% ± 0.2%	34.2% ± 1.7%	1.8% ± 0.3%
gay	13.8% ± 1.3%	4.0% ± 0.2%	34.3% ± 1.8%	2.4% ± 0.3%
location	16.3% ± 1.5%	4.8% ± 0.3%	25.3% ± 2.1%	2.4% ± 0.4%
payday	17.4% ± 1.4%	3.9% ± 0.2%	29.7% ± 1.7%	1.4% ± 0.3%
prostate	52.6% ± 6.8%	3.7% ± 0.3%	34.6% ± 1.4%	0.9% ± 0.2%
unemployed	14.3% ± 1.2%	4.5% ± 0.3%	22.8% ± 1.8%	2.9% ± 0.5%
other	17.8% ± 27.9%	3.7% ± 0.2%	27.5% ± 1.5%	1.4% ± 0.2%

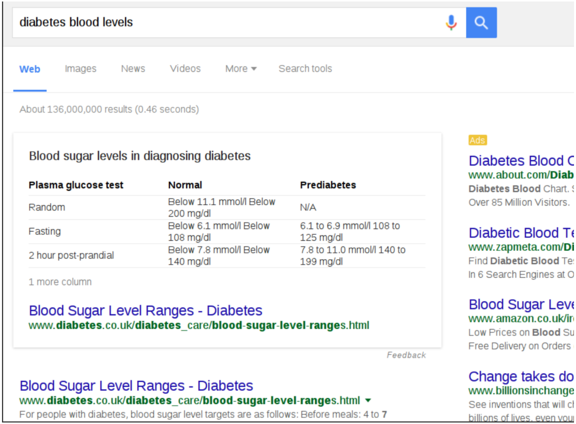
Average percentage content change per instance of probe query, grouped by topic and search engine

- Can we exploit the responsiveness of adverts ? Remember for-profit organisations are obliged to maximise shareholder value. So if they think they know information about us that will do that ...

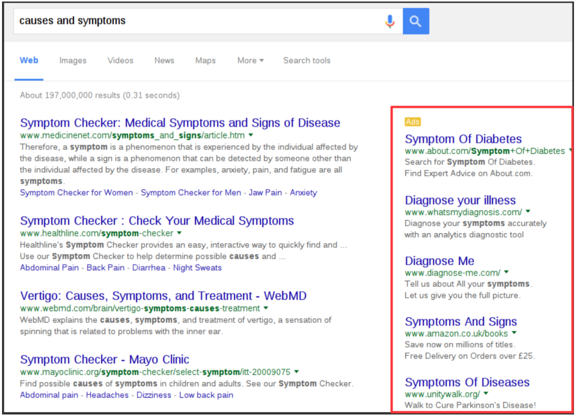
Online Search: Example of Using Probe Query



Online Search: Example of Using Probes Query



Online Search: Example of Using Probe Query



- Looking for clear evidence of learning → confident in our inferences.
- If we query "diabetes" and see adverts related to diabetes, then that is weaker evidence that if we query "causes and symptoms" repeatedly and see diabetes adverts start to appear over time.

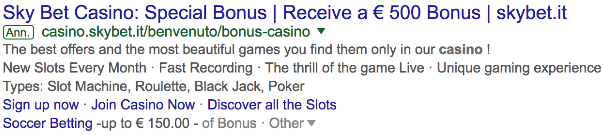
Details[3]

- Training data T. Consisting of displayed adverts tagged with a category e.g. sensitive/non-sensitive. Collect this data by a web crawl or logging adverts displayed to users via a browser extension.
- Construct dictionary D
  - Remove stop words e.g. a, of, the
  - Stemming e.g truncate {clicking, clicks, clicked} to click.
  - Assign unique integer id to resulting tokens from $T \to$ dictionary D
- Vectorization of adverts
  - Convert advert text to an integer vector x of size $|D|$ where value of x_w is the number of times token w appears in the advert (i.e. a "bag of words" model)

[3] Pol MacAonghusa, DL, *Don't Let Google Know I'm Lonely*, ACM Trans Privacy & Security 2016

Details

Example of Ad Vectorisation



- Suppose dictionary D=[roulette, gam,casino,bet]
- "roulette" appears once in ad
- "gam" appears three times ("games", "game", "gaming" are all stemmed to "gam").
- Vector associated with advert would be $x = [1, 3, 3, 2]$
- We do not include words not in dictionary D.
- Note that we can apply this process to create a vector associated with a single ad, to all of the adverts on a page etc

Details

- Using our labelled training data T ...
  - For each keyword calculate (i) what fraction of adverts on a particular topic c contain the keyword and (ii) what fraction of all adverts contain the keyword.
  - E.g. keyword "casino" might appear in 90% of online betting adverts but only in 5% of adverts overall. In which case its potentially a good indicator of the topic "gambling".
- Given a new page with adverts ...
  - We need to give more weight to keywords that appear frequently in the page
  - PRI estimator:

$$\sum_{w \in D} \left(\frac{\text{fraction of topic } c \text{ adverts with keyword } w}{\text{fraction of all adverts with keyword } w} \times \text{fraction of } w \text{ on page} \right)$$

  - Value ≈ 1 indicates no evidence of learning. Large or small values show evidence of learning – use threshold based on training data.

Detection Accuracy

- Measured data: 37,134 queries and responses collected over 28 days.
- Queries are associated with 12 topics:
  - 10 are associated with discrimination (health, disability, sexual orientation) or sensitive personal conditions (gambling addiction, financial problems)
  - 1 is location related
  - 1 consists of the top 50 queries from Google Trends.
- Sensitive topic keywords are taken from wikipedia and open directory project.
- Queries augmented with common words e.g. "fat" $\rightarrow$ why am i so fat"

Detection Accuracy

	Reference Topic										
	anorexia	bankrupt	diabetes	disabled	divorce	gambling	gay	location	payday	prostate	unemployed
True Detect	100%	100%	96%	100%	100%	100%	100%	99%	99%	99%	100%
True Other	96%	96%	92%	100%	100%	100%	100%	100%	100%	100%	100%
False Detect	4%	4%	8%	0%	0%	0%	0%	0%	0%	0%	0%
False Other	0%	0%	4%	0%	0%	0%	0%	1%	1%	1%	0%

Measured detection rate of google search learning of individual sensitive topics[4]

- True Detect = learning of topic detected, False Detect = mistake
- True Other = no evidence found of topic learning, False Other = mistake

[4]Pol MacAonghusa, DL, *Don't Let Google Know I'm Lonely*, ACM Trans Privacy & Security 2016

Speed of Learning By Search Engines

Number of Consecutive Misclassifications (X)			Probe ID of First Misclassification (Y)		
	Bing	**Google**		**Bing**	**Google**
$\mathbb{P}(X=1)$	0.23	0.95	$\mathbb{P}(Y=1)$	0.92	0.98
$\mathbb{P}(X=2)$	0.77	0.05	$\mathbb{P}(Y=2)$	0.03	0.01
$\mathbb{P}(X=3)$	0.00	0.00	$\mathbb{P}(Y=3)$	0.04	0.01
$\mathbb{P}(X=4)$	0.00	0.00	$\mathbb{P}(Y=4)$	0.00	0.00
$\mathbb{P}(X=5)$	0.00	0.00	$\mathbb{P}(Y=5)$	0.00	0.00

Probability of misclassification vs number of probes. $E[X|Google] = 1.05$, $E[X|Bing] = 1.77$

- Probes are sent on average every 4 queries, so Google appears to adapt to a new topic in approx 4 queries.

Logged In vs Anonymous

	Reference Topic										
	anorexia	**bankrupt**	**diabetes**	**disabled**	**divorce**	**gambling**	**gay**	**location**	**payday**	**prostate**	**unemployed**
True Detect	97%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
True Other	100%	100%	92%	100%	100%	100%	100%	100%	100%	100%	100%
False Detect	4%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%
False Other	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Measured detection rate of google search learning of individual sensitive topics when user is not logged in

Can We Slow Down/Disrupt Learning ?

Some ideas:

- Inject noisy queries into the user session (can be invisible to user if done in background via a browser extension). E.g.
  - Inject popular queries drawn from Google Trends
  - Be more sophisticated and try to mimic multiple user sessions
- User clicks can reveal user interests, so might also disrupt click behaviour. E.g.
  - Click nothing
  - Click all links
  - Click randomly selected links
  - Click links related to topics other than that of the user session

Do any of these strategies work ?

Plausible Deniability[5]

- Previous PRI detection approach estimates:

$$M_k(c) = \frac{P(\text{interest in topic } c | \text{user session up to } k\text{'th search})}{P(\text{interest in topic } c)}$$

- For a set of topics C and any pair of topics $c, d \in C$ for *plausible deniability* require:

$$e^{-\epsilon} \leq D_k(c,d) := \frac{P(\text{user session up to } k\text{'th search} \mid \text{interest in topic } c)}{P(\text{user session up to } k\text{'th search} \mid \text{interest in topic } d)} \leq e^{\epsilon}$$

- i.e. Require $|\log D_k(c,d)| \leq \epsilon$ for all c, d
- Using Bayes Rule:

$$D_k(c,d) = \frac{M_k(c)M_1(d)}{M_k(d)M_1(c)}$$

so can reuse PRI to estimate $D_k(c,d)$.

[5] Pol MacAonghusa, DL, *Plausible Deniability in Web Search*, IEEE Trans Inf Forensics & Security 2017

Measured Performance[6]

- Measured data: 21,861 queries and responses collected using same approach as before.
- For each test configuration responses for at least 2000 probe queries were collected, with 1000 used as training data and the other 1000 used for testing (with 7-fold cross-validation used to estimate sensitivity).

[6] Pol MacAonghusa, DL, *Plausible Deniability in Web Search*, IEEE Trans Inf Forensics & Security 2017

Measured Performance: Injecting Noise

Reference Topic	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5
anorexia	48 (48)	48 (48)	48 (48)	48 (48)	48 (48)
bankrupt	16 (10)	65 (51)	65 (48)	65 (49)	65 (49)
diabetes	41 (38)	41 (38)	41 (38)	41 (38)	41 (38)
disabled	9 (9)	9 (9)	9 (5)	9 (7)	9 (8)
divorce	41 (27)	75 (38)	56 (22)	75 (29)	75 (29)
gambling	21 (16)	21 (3)	21 (4)	29 (16)	18 (4)
gay	86 (64)	86 (64)	80 (43)	94 (59)	94 (59)
location	10 (10)	8 (8)	8 (8)	18 (13)	18 (13)
payday	3 (2)	4 (2)	4 (2)	4 (2)	3 (1)
prostate	17 (15)	17 (15)	17 (15)	17 (15)	17 (15)
unemployed	10 (7)	13 (7)	13 (7)	13 (7)	13 (7)

Estimated $|\log D_k(c,d)|$ with no clicks and 3 random queries inserted after every user query. Reported as "Max (Median)" percentages[7].

- $e^0 = 1$
- $e^{0.5} \approx 1.6$, $e^{-0.5} \approx 0.6$
- $e^{0.75} \approx 2.1$, $e^{-0.75} \approx 0.5$

[7] Pol MacAonghusa, DL, *Plausible Deniability in Web Search*, IEEE Trans Inf Forensics & Security 2017

Measured Performance: Random Clicks

Reference Topic	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5
anorexia	50 (12)	27 (9)	26 (9)	36 (10)	33 (11)
bankrupt	5 (3)	43 (33)	39 (37)	36 (35)	38 (35)
diabetes	38 (6)	18 (7)	17 (5)	17 (7)	11 (5)
disabled	2 (1)	4 (1)	5 (3)	39 (25)	40 (25)
divorce	24 (17)	37 (31)	37 (31)	35 (25)	35 (25)
gambling	24 (0)	7 (4)	54 (23)	33 (23)	68 (20)
gay	68 (68)	68 (65)	54 (52)	46 (36)	47 (42)
location	8 (8)	8 (8)	8 (8)	8 (8)	8 (8)
payday	4 (1)	2 (2)	4 (2)	4 (3)	4 (4)
prostate	59 (57)	67 (62)	58 (56)	60 (54)	51 (44)
unemployed	4 (3)	8 (3)	10 (4)	3 (2)	10 (1)

Estimated $|\log D_k(c,d)|$ when click 2 links at random.

- $e^0 = 1$
- $e^{0.5} \approx 1.6$, $e^{-0.5} \approx 0.6$
- $e^{0.75} \approx 2.1$, $e^{-0.75} \approx 0.5$

Measured Performance: Clicking All

Reference Topic	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5
anorexia	66 (57)	66 (57)	66 (57)	66 (57)	66 (57)
bankrupt	51 (42)	51 (42)	51 (42)	55 (46)	56 (46)
diabetes	35 (35)	35 (35)	35 (35)	35 (35)	35 (35)
disabled	9 (9)	9 (9)	9 (9)	31 (31)	31 (31)
divorce	30 (8)	73 (54)	54 (34)	100 (49)	100 (49)
gambling	3 (1)	16 (16)	53 (11)	16 (6)	6 (2)
gay	69 (65)	77 (73)	70 (60)	82 (75)	81 (71)
location	18 (10)	10 (6)	10 (6)	14 (10)	18 (7)
payday	2 (2)	2 (2)	2 (2)	2 (2)	2 (2)
prostate	17 (17)	17 (17)	17 (17)	17 (17)	17 (17)
unemployed	4 (4)	7 (7)	7 (7)	7 (7)	7 (6)

Estimated $|\log D_k(c, d)|$ when click all links.

- $e^0 = 1$
- $e^{0.5} \approx 1.6$, $e^{-0.5} \approx 0.6$
- $e^{0.75} \approx 2.1$, $e^{-0.75} \approx 0.5$

Measured Performance: Injecting Fake User Sessions

- Inject sequence of related queries. Topics used: tickets for events in Croke Park, vacation flights/hotels, car trade-in.
- Now have two interleaved sessions, the true user session and a fake one. Inject 3-4 fake queries for every true query, and randomly shuffle.

Reference Topic	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5
all topics	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

Estimated $|\log D_k(c,d)|$. Zero for all topics and all click strategies.

- $e^0 = 1$

Conclusions

- Its fairly straightforward to provide users with feedback on observed learning
- Learning by Google seems to fast, accurate and robust
- We can effectively disrupt this learning by injecting fake user sessions.
- But it seems like an arms race – personal queries are inherently revealing and we can expect use of more sophisticated learning approaches by the search engine to be able to extract them e.g. use of mixture models.
- Are there alternatives to this arms race ? Perhaps hiding in the crowd/clustering is an option – if were interested in personalised adverts, then we can likely get state of the art performance with much less personal information[8].

[8] A.Checco, G.Bianchi, DL, *BLC: Private Matrix Factorization Recommenders via Automatic Group Learning*, ACM Trans Privacy & Security 2017