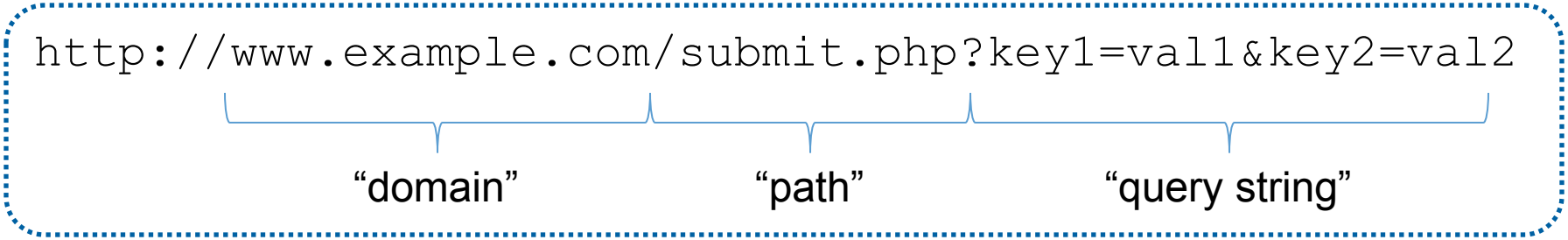
On the Privacy Concerns of URL Query Strings

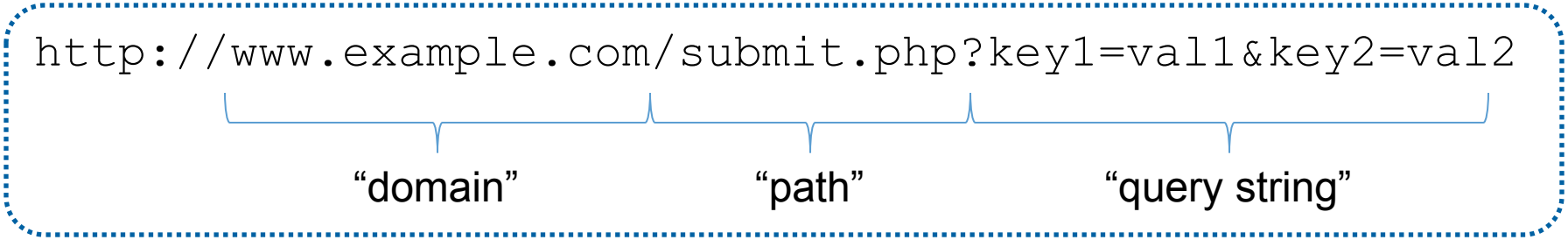
Andrew G. West (Verisign Labs) and Adam J. Aviv (USNA)

May 18, 2014 – Web 2.0 Security & Privacy

URL Query Strings

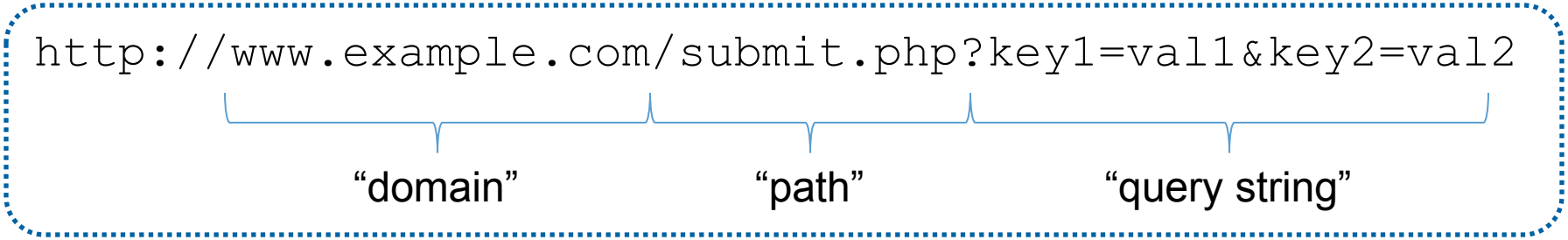


URL Query Strings



- Server-side languages: ASP, CGI, JS, PHP
- 56% of URLs (in our data) have 1+ key-value pairs

URL Query Strings



- Server-side languages: ASP, CGI, JS, PHP
- 56% of URLs (in our data) have 1+ key-value pairs
- Primarily opaque IDs; sometimes privacy-sensitive
- Exacerbated by Web 2.0 social services; info sharing culture

URL Query Strings

`http://www.example.com/submit.php?key1=val1&key2=val2`

"domain" "path" "query string"

Copy-pasted URLS

- Server-side languages: ASP, CGI, JS, PHP
- 56% of URLs (in our data) have 1+ key-value pairs
- Primarily opaque IDs; sometimes privacy-sensitive
- Exacerbated by Web 2.0 social services; info sharing culture

URL Query Strings

`http://www.example.com/submit.php?key1=val1&key2=val2`

"domain" "path" "query string"

Copy-pasted URLS

- Server-side languages: ASP, CGI, JS, PHP
- 56% of URLs (in our data) have 1+ key-value pairs
- Primarily opaque IDs; sometimes privacy-sensitive
- Exacerbated by Web 2.0 social services; info sharing culture

URL Query Strings

`http://www.example.com/submit.php?key1=val1&key2=val2`

"domain" "path" "query string"

Copy-pasted URLS

- Server-side languages: ASP, CGI, JS, PHP
- 56% of URLs (in our data) have 1+ key-value pairs
- Primarily opaque IDs; sometimes privacy-sensitive
- Exacerbated by Web 2.0 social services; info sharing culture

The Authors' Position

URL-BASED PRIVACY CONCERNS ARE SIGNIFICANT

- In 892M URLs in public domain we find:

The Authors' Position

URL-BASED PRIVACY CONCERNS ARE SIGNIFICANT

- In 892M URLs in public domain we find:
  - Quarter *billion* instances of referral data

The Authors' Position

URL-BASED PRIVACY CONCERNS ARE SIGNIFICANT

- In 892M URLs in public domain we find:
  - Quarter *billion* instances of referral data
  - 10+ million more sensitive fields (geo-location, network properties, online and physical identity, phone numbers, *etc.*)

The Authors' Position

URL-BASED PRIVACY CONCERNS ARE SIGNIFICANT

- In 892M URLs in public domain we find:
  - Quarter *billion* instances of referral data
  - 10+ million more sensitive fields (geo-location, network properties, online and physical identity, phone numbers, *etc.*)
  - Isolated examples of authentication tokens

The Authors' Position

URL-BASED PRIVACY CONCERNS ARE SIGNIFICANT

- In 892M URLs in public domain we find:
  - Quarter *billion* instances of referral data
  - 10+ million more sensitive fields (geo-location, network properties, online and physical identity, phone numbers, *etc.*)
  - Isolated examples of authentication tokens
- Non-intentional disclosures revealed in plain-text

The Authors' Position

URL-BASED PRIVACY CONCERNS ARE SIGNIFICANT

- In 892M URLs in public domain we find:
  - Quarter *billion* instances of referral data
  - 10+ million more sensitive fields (geo-location, network properties, online and physical identity, phone numbers, *etc.*)
  - Isolated examples of authentication tokens
- Non-intentional disclosures revealed in plain-text

WEB 2.0 SERVICES IDEAL FOR PRIVACY LOGIC

- Web 2.0 is medium by which many links arrive on public web
- Strip params unnecessary for rendering; retroactively sanitize

How do we approach this?

1. Measurement study over 892M user-sourced URLs
2. “CleanURL” (a privacy-aware link transformation service)

URL Corpus (Basic Properties)

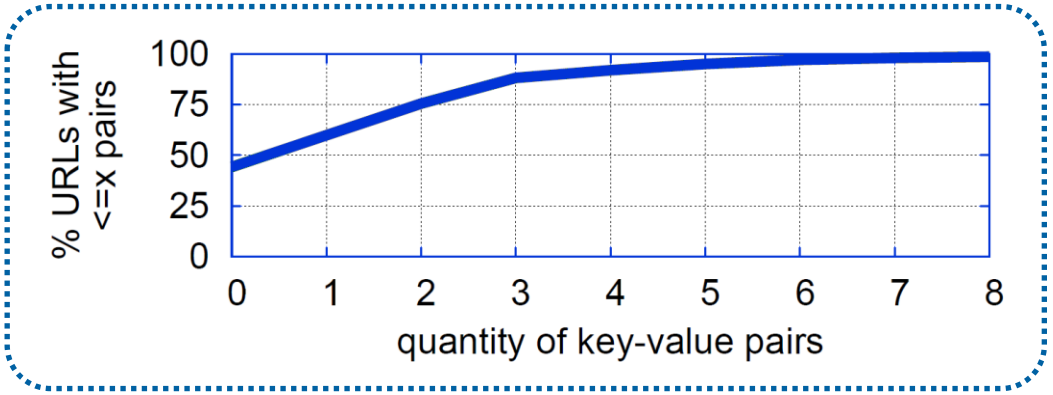
- ≈892 million URLs from early 2014
- Provided by an industry service provider

URL Corpus (Basic Properties)

- ≈892 million URLs from early 2014
- Provided by an industry service provider
- URLs submitted by end-users; provider's service eases link tracking and handling
- Links commonly found posted to Web 2.0 social services.

URL Corpus (Basic Properties)

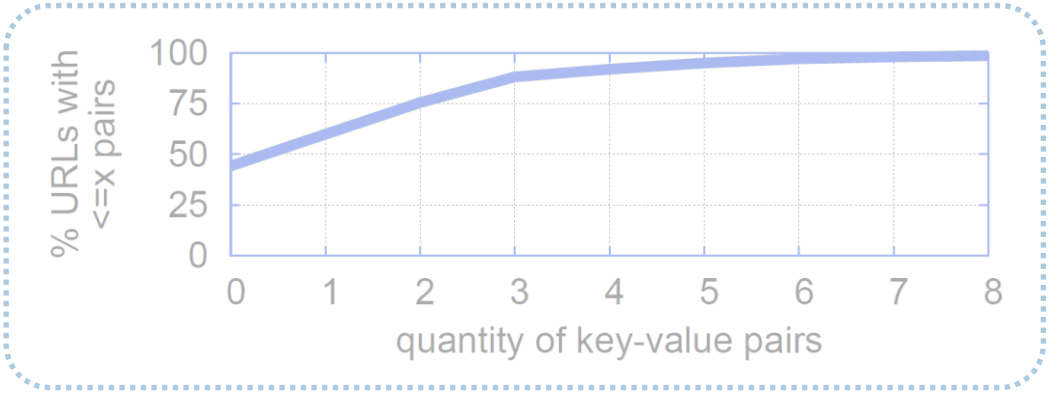
- ≈892 million URLs from early 2014
- Provided by an industry service provider
- URLs submitted by end-users; provider's service eases link tracking and handling
- Links commonly found posted to Web 2.0 social services.



- How common are parameters:
  - 490M URLs (54.9%) w/1+ pair
  - 44.6M URLs (5%) w/5+ pairs
  - 23.4K URLS w/100+ pairs

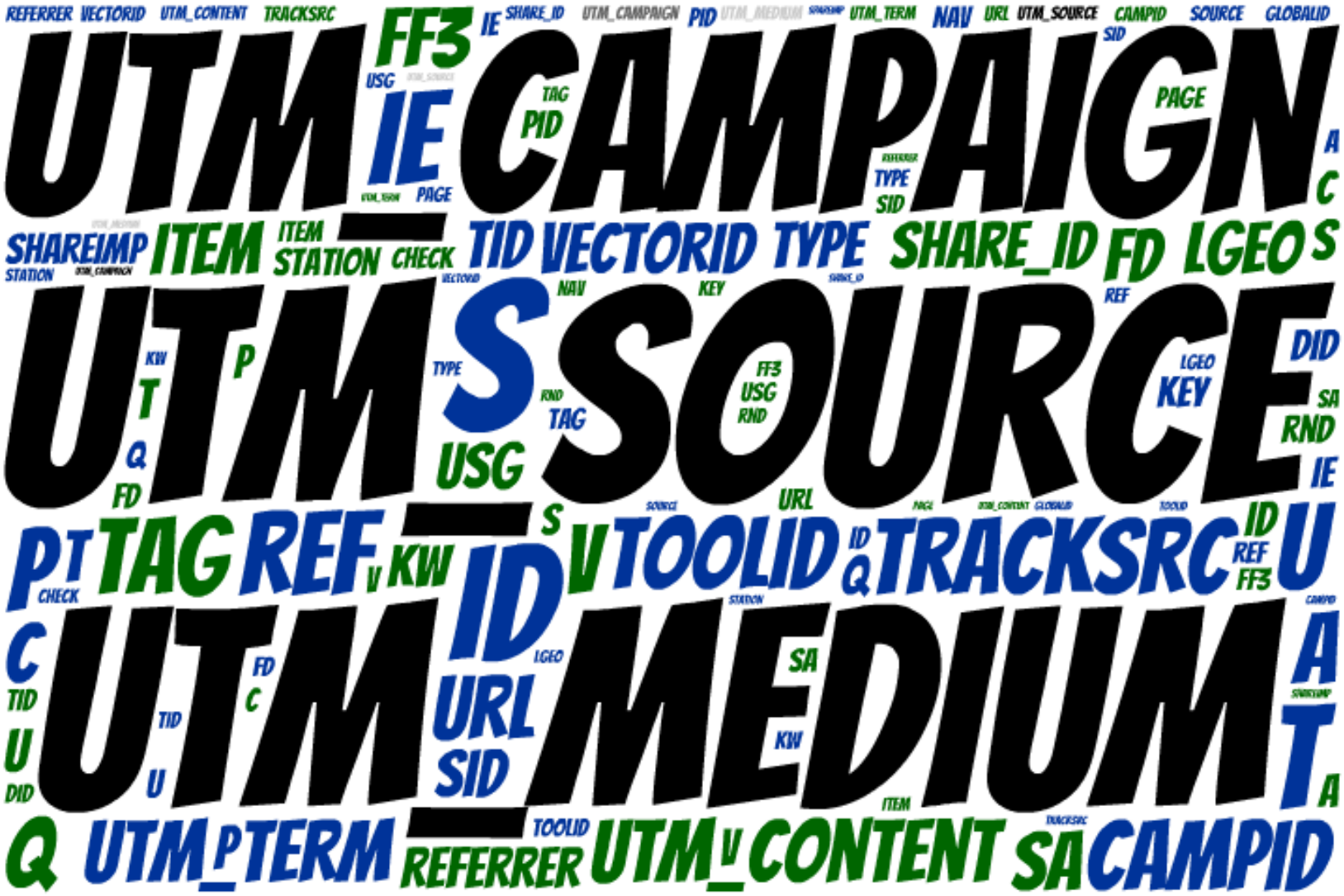
URL Corpus (Basic Properties)

- ≈892 million URLs from early 2014
- Provided by an industry service provider
- URLs submitted by end-users; provider's service eases link tracking and handling
- Links commonly found posted to Web 2.0 social services.



- How common are parameters:
  - 490M URLs (54.9%) w/1+ pair
  - 44.6M URLs (5%) w/5+ pairs
  - 23.4K URLS w/100+ pairs
- Broader perspective:
  - 1.3 billion key-value pairs total
  - 909k unique key names

Common Query String Keys



Privacy-sensitive Key-value Pairs

THEME	KEYS	SUM#
ALL URLs	----	892,934,790
URLs w/keys	****	490,227,789

* Produced using Monte-Carlo over manual inspection of 861 keys w/100k+ instances

Privacy-sensitive Key-value Pairs

THEME	KEYS	SUM#
ALL URLs URLs w/keys	---- ****	892,934,790 490,227,789
Referrer data	utm_source, ref, tracksrc, referrer, source, src, sentFrom, referralSource, referral_source	259,490,318

* Produced using Monte-Carlo over manual inspection of 861 keys w/100k+ instances

Privacy-sensitive Key-value Pairs

THEME	KEYS	SUM#
ALL URLs URLs w/keys	---- ****	892,934,790 490,227,789
Referrer data	utm_source, ref, tracksrc, referrer, source, src, sentFrom, referralSource, referral_source	259,490,318
Geo-location	my_lat, my_lon, zip, country, coordinate, hours_offset, address	5,961,565

* Produced using Monte-Carlo over manual inspection of 861 keys w/100k+ instances

Privacy-sensitive Key-value Pairs

THEME	KEYS	SUM#
ALL URLs URLs w/keys	---- ****	892,934,790 490,227,789
Referrer data	utm_source, ref, tracksrc, referrer, source, src, sentFrom, referralSource, referral_source	259,490,318
Geo-location	my_lat, my_lon, zip, country, coordinate, hours_offset, address	5,961,565
Network	ul_speed, dl_speed, network_name, mobile	3,824,398

* Produced using Monte-Carlo over manual inspection of 861 keys w/100k+ instances

Privacy-sensitive Key-value Pairs

THEME	KEYS	SUM#
ALL URLs URLs w/keys	---- ****	892,934,790 490,227,789
Referrer data	utm_source, ref, tracksrc, referrer, source, src, sentFrom, referralSource, referral_source	259,490,318
Geo-location	my_lat, my_lon, zip, country, coordinate, hours_offset, address	5,961,565
Network	ul_speed, dl_speed, network_name, mobile	3,824,398
Identity (online)	uname, user_email, email, user_id, user, login_account_id	2,142,654

* Produced using Monte-Carlo over manual inspection of 861 keys w/100k+ instances

Privacy-sensitive Key-value Pairs

THEME	KEYS	SUM#
ALL URLs URLs w/keys	---- ****	892,934,790 490,227,789
Referrer data	utm_source, ref, tracksrc, referrer, source, src, sentFrom, referralSource, referral_source	259,490,318
Geo-location	my_lat, my_lon, zip, country, coordinate, hours_offset, address	5,961,565
Network	ul_speed, dl_speed, network_name, mobile	3,824,398
Identity (online)	uname, user_email, email, user_id, user, login_account_id	2,142,654
Authentication	login_password, pwd	672,948

* Produced using Monte-Carlo over manual inspection of 861 keys w/100k+ instances

Privacy-sensitive Key-value Pairs

THEME	KEYS	SUM#
ALL URLs URLs w/keys	---- ****	892,934,790 490,227,789
Referrer data	utm_source, ref, tracksrc, referrer, source, src, sentFrom, referralSource, referral_source	259,490,318
Geo-location	my_lat, my_lon, zip, country, coordinate, hours_offset, address	5,961,565
Network	ul_speed, dl_speed, network_name, mobile	3,824,398
Identity (online)	uname, user_email, email, user_id, user, login_account_id	2,142,654
Authentication	login_password, pwd	672,948
Identity (real)	name1, name2, gender	533,222

* Produced using Monte-Carlo over manual inspection of 861 keys w/100k+ instances

Privacy-sensitive Key-value Pairs

THEME	KEYS	SUM#
ALL URLs URLs w/keys	---- ****	892,934,790 490,227,789
Referrer data	utm_source, ref, tracksrc, referrer, source, src, sentFrom, referralSource, referral_source	259,490,318
Geo-location	my_lat, my_lon, zip, country, coordinate, hours_offset, address	5,961,565
Network	ul_speed, dl_speed, network_name, mobile	3,824,398
Identity (online)	uname, user_email, email, user_id, user, login_account_id	2,142,654
Authentication	login_password, pwd	672,948
Identity (real)	name1, name2, gender	533,222
Phone	phone	56,267

* Produced using Monte-Carlo over manual inspection of 861 keys w/100k+ instances

Privacy-sensitive Key-value Pairs (2)

THEME	KEYS	SUM#
ALL		790
UR		789
Ref		318
Ge		565
Network	ul_speed, dl_speed, network_name, mobile	3,824,398
Identity (online)	uname, user_email, email, user_id, user, login_account_id	2,142,654
Authentication	login_password, pwd	672,948
Identity (real)	name1, name2, gender	533,222
Phone	phone	56,267

Prevalence may be *under-reported*

- Naming conventions are non-standardized:
  - 103K instances of key "`email`"
  - 637K (6.2×) keys pattern match "`*email*`"
  - 1.7M (16.5×) instances where value is an email address
  - 2000+ unique keys have email values

* Produced using Monte-Carlo over manual inspection of 861 keys w/100k+ instances

Privacy-sensitive Key-value Pairs (2)

THEME	KEYS	SUM#
AL		790
UR		789
Re		318
Ge		565
Network	ul_speed, dl_speed,	3,824,398
Ide		654
Au		948
Ide		222
Phone	phone	56,267

Prevalence may be *under-reported*

- Naming conventions are non-standardized:
  - 103K instances of key "`email`"
  - 637K (6.2×) keys pattern match "`*email*`"
  - 1.7M (16.5×) instances where value is an email address
  - 2000+ unique keys have email values

Must be *cautious* of such claims

- Not all values are sensitive (just a majority per Monte Carlo)
- No idea which of these values are "personal"
  - Ex: do geo-coordinates locate user? Or a monument?

* Produced using Monte-Carlo over manual inspection of 861 keys w/100k+ instances

Authentication Tokens in Query Strings

- Password values are *almost* always encrypted

Authentication Tokens in Query Strings

- Password values are *almost* always encrypted
- Best practices adhered to (*i.e.*, salting)
  - Variable-length MD5/SHA hashes of 100 most common passwords produced no hits in our corpus

Authentication Tokens in Query Strings

- Password values are *almost* always encrypted
- Best practices adhered to (*i.e.*, salting)
  - Variable-length MD5/SHA hashes of 100 most common passwords produced no hits in our corpus
- Several dozen instances of full credentials in plain-text

Authentication Tokens in Query Strings

- Password values are *almost* always encrypted
- Best practices adhered to (*i.e.*, salting)
  - Variable-length MD5/SHA hashes of 100 most common passwords produced no hits in our corpus
- Several dozen instances of full credentials in plain-text

“Grand slam” examples, redacted:

```
[media]/xmlrpc.php?cmd=getVideos&username=admin&password=[redacted]

[medical]/index.aspx?accountname=[redacted]health&username=[redacted]&password=[redacted]

[healthcare]/?do=patient&directAccess=yes&username=[redacted]&password=[redacted]
```

Value Entropy

- Diversity/entropy of key's value set
  - Few values = little diversity = less revealing (*e.g.*, gender)

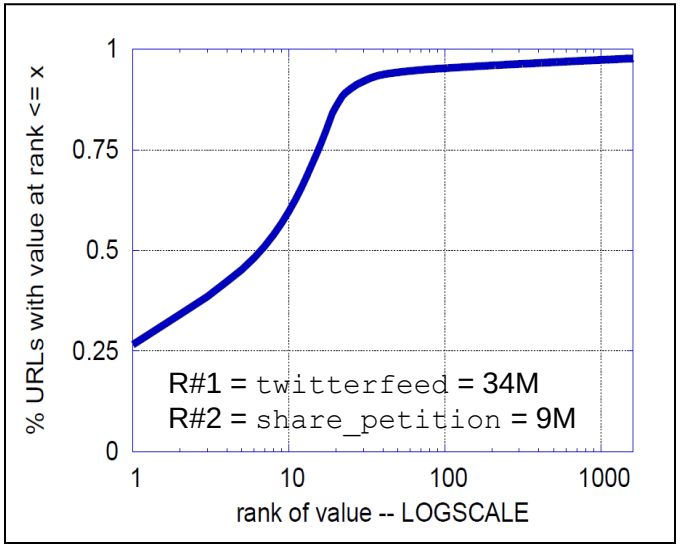
Value Entropy

- Diversity/entropy of key's value set
  - Few values = little diversity = less revealing (*e.g.*, gender)
  - Diversity calculation, d , lies on [0,1]
  - Most privacy-relevant keys on $0.33 < d < 0.66$

Value Entropy

- Diversity/entropy of key's value set
  - Few values = little diversity = less revealing (*e.g.*, gender)
  - Diversity calculation, *d* , lies on [0,1]
  - Most privacy-relevant keys on $0.33 < d < 0.66$
- Distribution of value set also interesting:

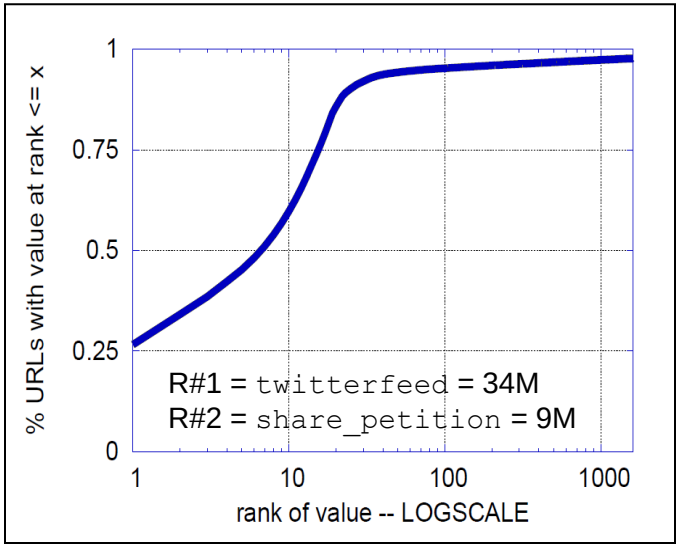
key = `utm_source` (128M instances)



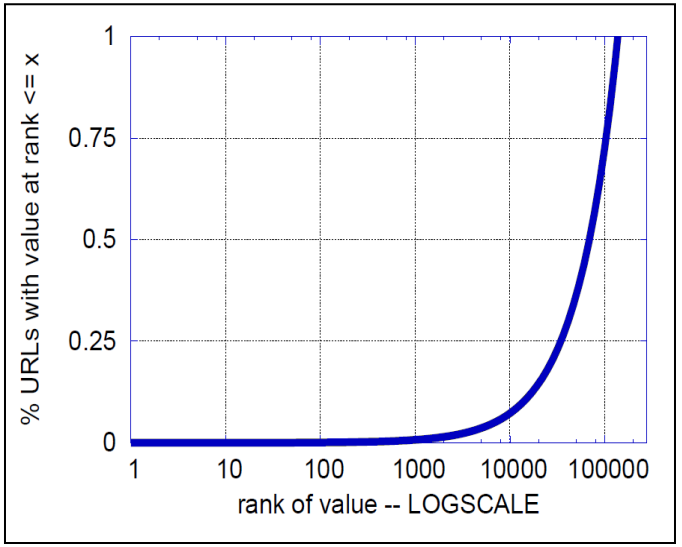
Value Entropy

- Diversity/entropy of key's value set
  - Few values = little diversity = less revealing (*e.g.*, gender)
  - Diversity calculation, *d* , lies on [0,1]
  - Most privacy-relevant keys on $0.33 < d < 0.66$
- Distribution of value set also interesting:

key = `utm_source` (128M instances)



key = `secureCode` (275k instances)



How do we approach this?

1. Measurement study over 892M user-sourced URLs
2. “CleanURL” (a privacy-aware link transformation service)

Argument Removal Logic

Key-value *NECESSITY*

- Is pair needed for faithful rendering?

Argument Removal Logic

Key-value *NECESSITY*

- Is pair needed for faithful rendering?

(1) No change w/removal

`zip` = 12345 (remove)

Argument Removal Logic

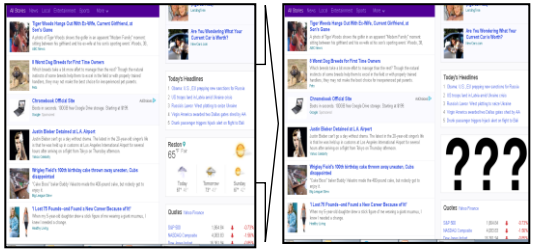
Key-value *NECESSITY*

- Is pair needed for faithful rendering?

(1) No change w/removal

`zip` = 12345 (remove)

(2) Orthogonal to main content



`zip` = 12345 (remove)

Argument Removal Logic

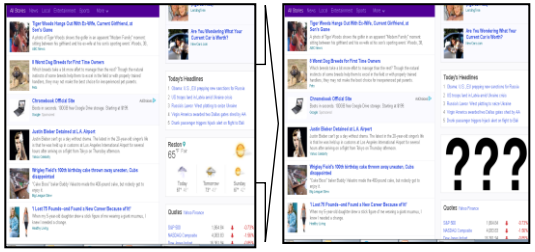
Key-value *NECESSITY*

- Is pair needed for faithful rendering?

(1) No change w/removal

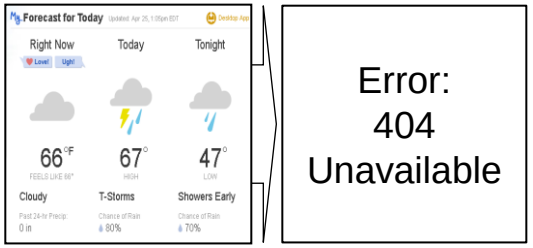
`zip` = 12345 (remove)

(2) Orthogonal to main content



`zip` = 12345 (remove)

(3) Unfaithful render



`zip` = 12345 (warn user)

Argument Removal Logic

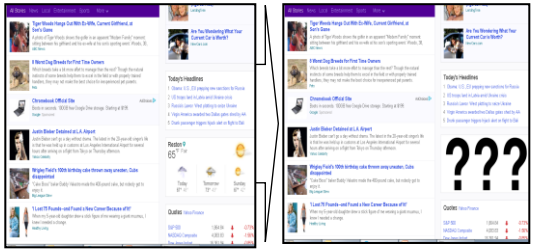
Key-value *NECESSITY*

- Is pair needed for faithful rendering?
- Programmatically difficult
  - Visual hamming distance
  - HTML tag delta size

(1) No change w/removal

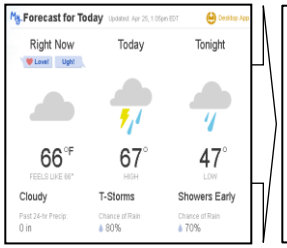
`zip` = 12345 (remove)

(2) Orthogonal to main content



`zip` = 12345 (remove)

(3) Unfaithful render



Error:
404
Unavailable

`zip` = 12345 (warn user)

Argument Removal Logic

Key-value *NECESSITY*

- Is pair needed for faithful rendering?
- Programmatically difficult
  - Visual hamming distance
  - HTML tag delta size

Key-value *SENSITIVITY*

- Does pair contain private information?

(1) No change w/removal

`zip` = 12345 (remove)

(2) Orthogonal to main content

`zip` = 12345 (remove)

(3) Unfaithful render

`zip` = 12345 (warn user)

Argument Removal Logic

Key-value *NECESSITY*

- Is pair needed for faithful rendering?
- Programmatically difficult
  - Visual hamming distance
  - HTML tag delta size

Key-value *SENSITIVITY*

- Does pair contain private information?
- Programmatically difficult
  - Regexes gleaned from manual work
  - Mining corpora w/metrics such as entropy
  - Human feedback loops once online

(1) No change w/removal

`zip` = 12345 (remove)

(2) Orthogonal to main content

`zip` = 12345 (remove)

(3) Unfaithful render

`zip` = 12345 (warn user)

CleanURL – Privacy Aware Link Transformer

CleanURL – Privacy Aware Link Transformer

❶

clean.URL
Because clean is sexy.

http://www.example.com?key1=val1... Submit

❷

Choose the left-most version that appears as you expect.
Our best guess has been selected by default.

www.example.com?~~key1=val1~~&key2=val2&~~key3=val3~~

CleanURL – Privacy Aware Link Transformer

❶

clean.URL
Because clean is sexy.

http://www.example.com?key1=val1... Submit

❷

Choose the left-most version that appears as you expect.
Our best guess has been selected by default.

www.example.com?~~key1=val1~~&key2=val2&~~key3=val3~~

❸

Your cleaned URL: [[base_url]]/R09XVIUh

Conclusion

POSITION: URL query strings have significant privacy impacts; social platforms should help curb issue as they are appropriate locales for privacy-preserving logic

Conclusion

POSITION: URL query strings have significant privacy impacts; social platforms should help curb issue as they are appropriate locales for privacy-preserving logic

- Motivational measurements over large URL corpus show personal data frequent and in plaintext
- CleanURL: A service proposed for URL sanitization

Conclusion

POSITION: URL query strings have significant privacy impacts; social platforms should help curb issue as they are appropriate locales for privacy-preserving logic

- Motivational measurements over large URL corpus show personal data frequent and in plaintext
- CleanURL: A service proposed for URL sanitization

CLOSING THOUGHTS / FUTURE:

- Direct scrapes off of the firehose/sprinkler APIs

Conclusion

POSITION: URL query strings have significant privacy impacts; social platforms should help curb issue as they are appropriate locales for privacy-preserving logic

- Motivational measurements over large URL corpus show personal data frequent and in plaintext
- CleanURL: A service proposed for URL sanitization

CLOSING THOUGHTS / FUTURE:

- Direct scrapes off of the firehose/sprinkler APIs
- Can domain sensitivity be learned from human feedback?
- Best practices involve HTTPS/TLS/SSL

powered by

VERISIGN™