COMPUTER SCIENCE
& ENGINEERING
TEXAS A&M UNIVERSITY

SocialSpamGuard

Dr. Martin “Doc” Carlisle

Premise

- Find spam social posts (unwanted, irrelevant, promotional, harmful)

SocialSpamGuard

- Scalable online social media spam detection
  - Automatically harvests spam activities
  - Utilize both image and text content
  - Clustering algorithm

Social Media Network Model

- Vertices = Users, Pages, Posts, Friendships/Followings, Fan/Favorites
- Edges = friendships/follows
- (content-similarity)
- Time-stamped

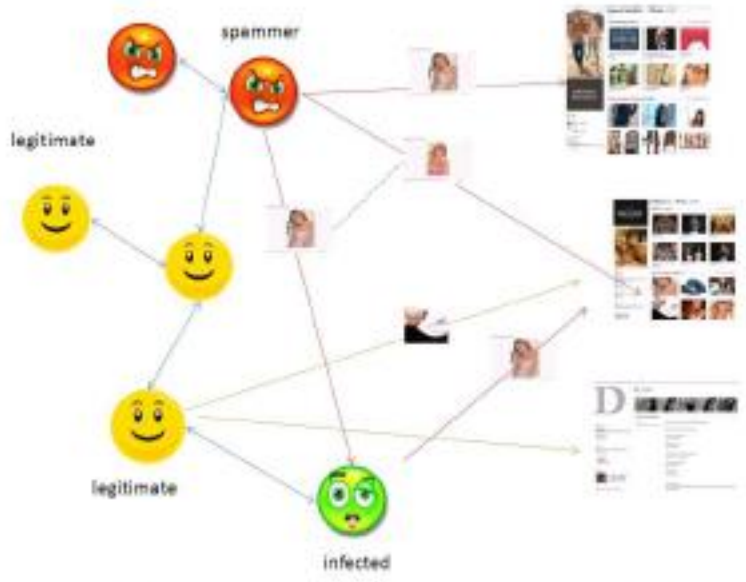


Figure 1: Heterogeneous Information Network for Social Media. A red face is a spammer, a yellow smile face is a legitimate user, a yellow face turned to green color is an infected user. The blue directed line is the friendship/following link. A red arrow is a spam post, while a green arrow is a ham post.

System Architecture

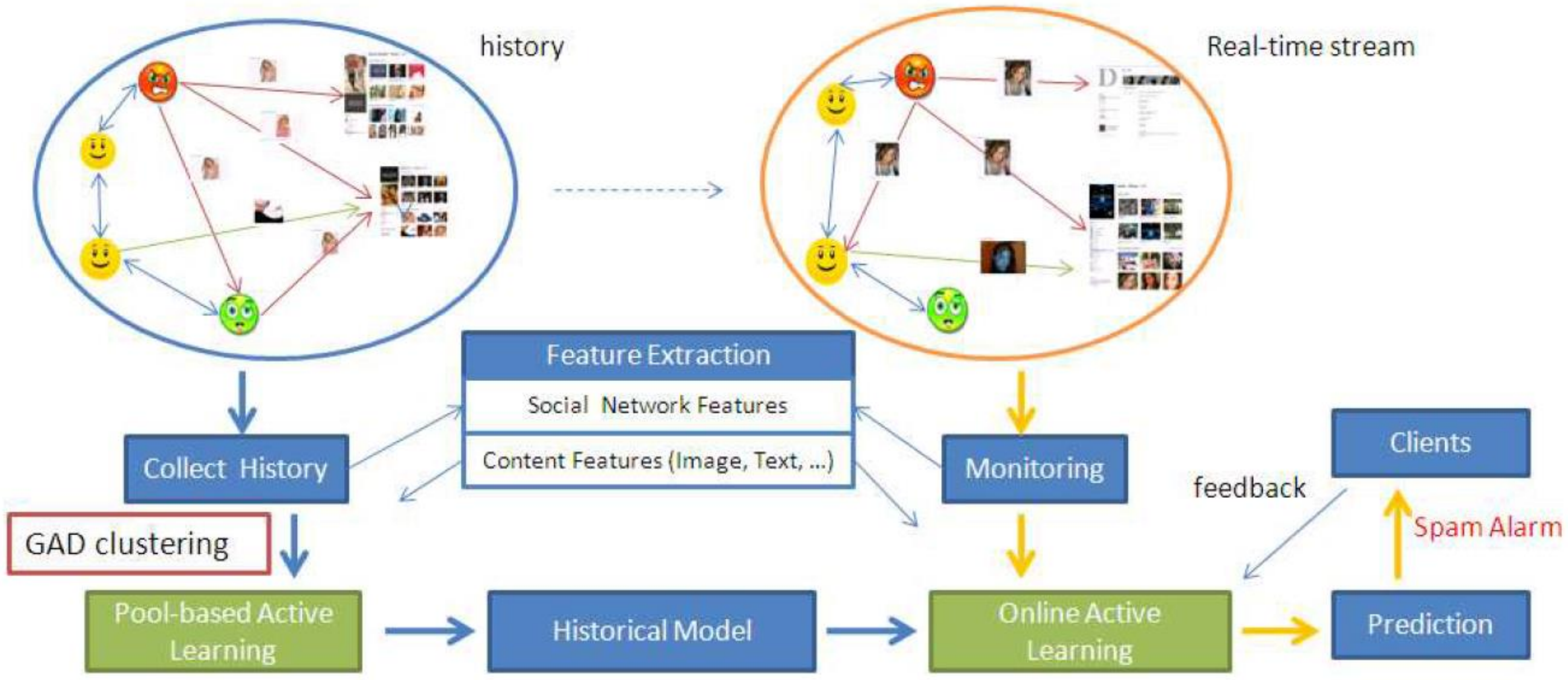


Figure 2: System Architecture.

Feature Content Extraction

- Image features
  - Color histogram
  - Color correlogram
  - Gabor features (texture analysis)
  - Edge histogram
  - SIFT (scale-invariant feature transform)
  - CEDD (color and edge directivity descriptor)
    - Next slide

CEDD

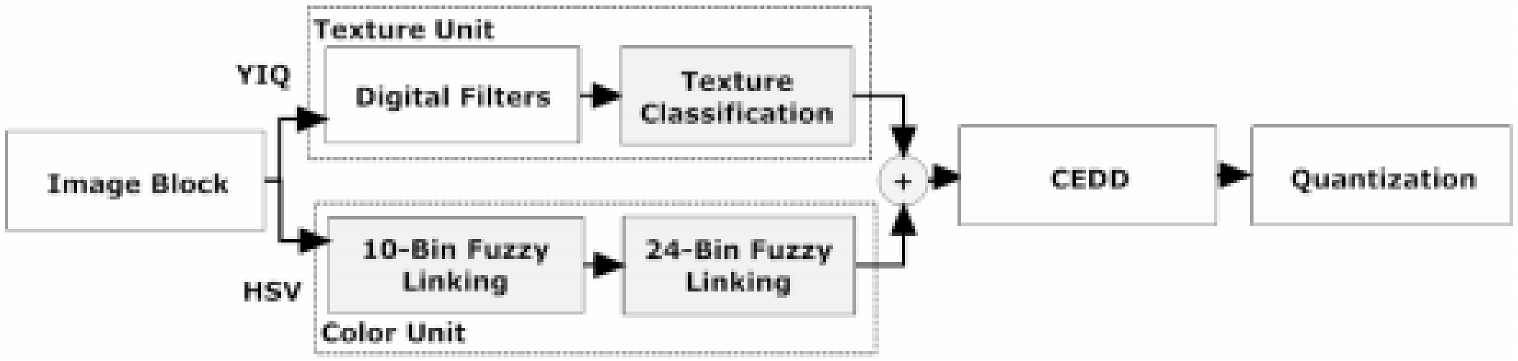


Fig. 4. CEDD Flowchart

Chatzichristofis S.A., Boutalis Y.S. (2008) CEDD: Color and Edge Directivity Descriptor: A Compact Descriptor for Image Indexing and Retrieval. In: Gasteratos A., Vincze M., Tsotsos J.K. (eds) Computer Vision Systems. ICVS 2008. Lecture Notes in Computer Science, vol 5008. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-79547-6_30

Feature Content Extraction

- Text features
  - Ratio of non-English words
  - Number of comments/likes
  - Number of sensitive words
  - Reputation of comment authors
  - Short URL leading to spam site (e.g. http://nxy.in/xxhpl)

Feature Content Extraction

- Social network features
  - Characteristics of profiles
  - Behaviors in network
    - Spammers don’t reply to comments (almost never)
    - Spammers post to popular pages
    - Spammers register as beautiful females/use celebrity names/photos
    - Often post similar to lots of pages

Scalable Active Learning for Historical Data

1. Generate initial set of instances for labeling, build classifier
2. Predict and rank remaining unlabeled (sort test posts in decreasing order & divide into blocks)
3. Obtain additional set of labeled posts (examine top blocks)
4. Add new labeled set to training pool and update model
5. Repeat 2-5 until stop criteria

GAD Clustering

- Random sampling may not be best
- Cluster posts into large number of clusters and sample from clusters to increase diversity

Online Active Learning

- Predict via trained model
  - Uncertain send for human labeling
  - When enough new labels, retrain

Case Study

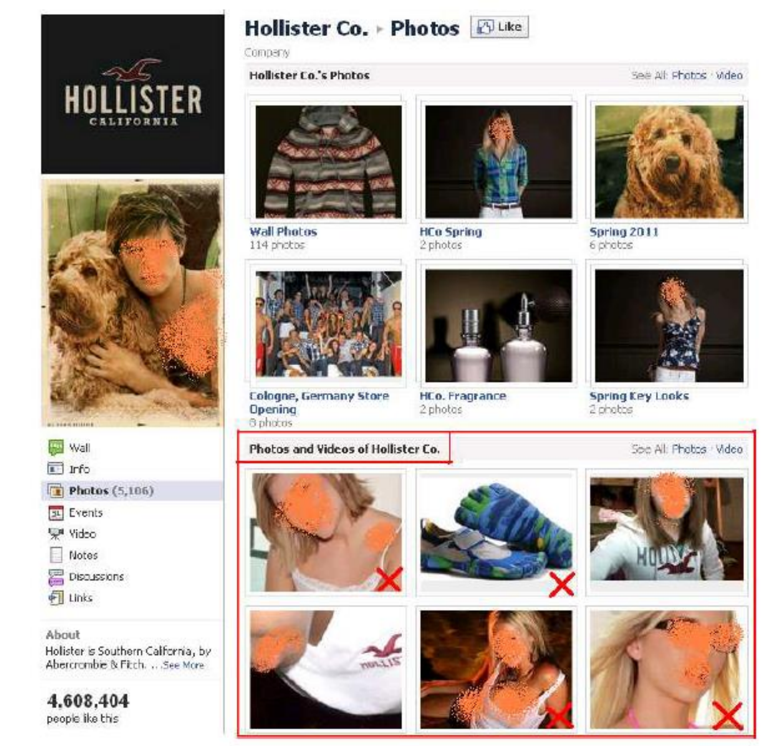


Figure 3: The Hollister Co. page on Facebook, accessed on March 28, 2011. The section "Photos and Videos of Hollister Co." (marked as red rectangle) lists the user added photos/videos in time decreasing order. Among the top 6 most recent photos, 4 of which are detected as spams (marked as red X). For privacy consideration, we have mosaicked the photos.

March 28, 2011- 4M fans,
5100 user added photos/videos

Top 6 recently added, 4 detected as spam.

First: "I am a very sweet woman and I am seeking for a gorgeous man to share a joy night with. See how gorgeous I am at http://nxy.in/xxhp1".

COMPUTER SCIENCE & ENGINEERING
TEXAS A&M UNIVERSITY

Detecting Bystanders in Photos

Dr. Martin “Doc” Carlisle

Premise

- Find bystanders in social media photos to improve privacy

What is a bystander?

- Someone who is “present but not taking part” in the photo
- Someone who is “not a subject of the photo and is thus not important for the meaning of the photo”

Other techniques

- Prevent image capture if bystander present
- Have bystanders broadcast a privacy policy
- Cloud solutions – users mark location private, or indicate to social network they want to be private

Dataset

- 91,118 images of 1-5 people from Google open image dataset (9.2M images)
- Randomly sampled 1307 (1 person), 615, 318, 206 and 137 (5 people) images, totaling 2,583 images. This corresponds to 5,000 faces.

Example Images

(a) Image with a single person. (b) Image with five people where the stimulus is enclosed by a bounding box. (c) An image where the annotated area contains a sculpture.

Fig. 1. Example stimuli used in our survey.

Survey questions

- Kind of image (person, depiction of person, something else)
- Public, semi-public, semi-private, private place
- Aware being photographed (1-7 Likert)
- Actively posing (1-7 Likert)
- Comfortable being photographed (1-7 Likert)
- Willing to be in photo (1-7)
- Can be replaced with random person w/o effect (1-7)
- Subject or bystander? Why?

Mechanical Turk

- Amazon micro-task service
- Restricted to USA at least 5 years, >=18 years old, with high reputation
- Paid $7 for about 41 minutes of work
- 387 people
  - Each image had at least 3 participants

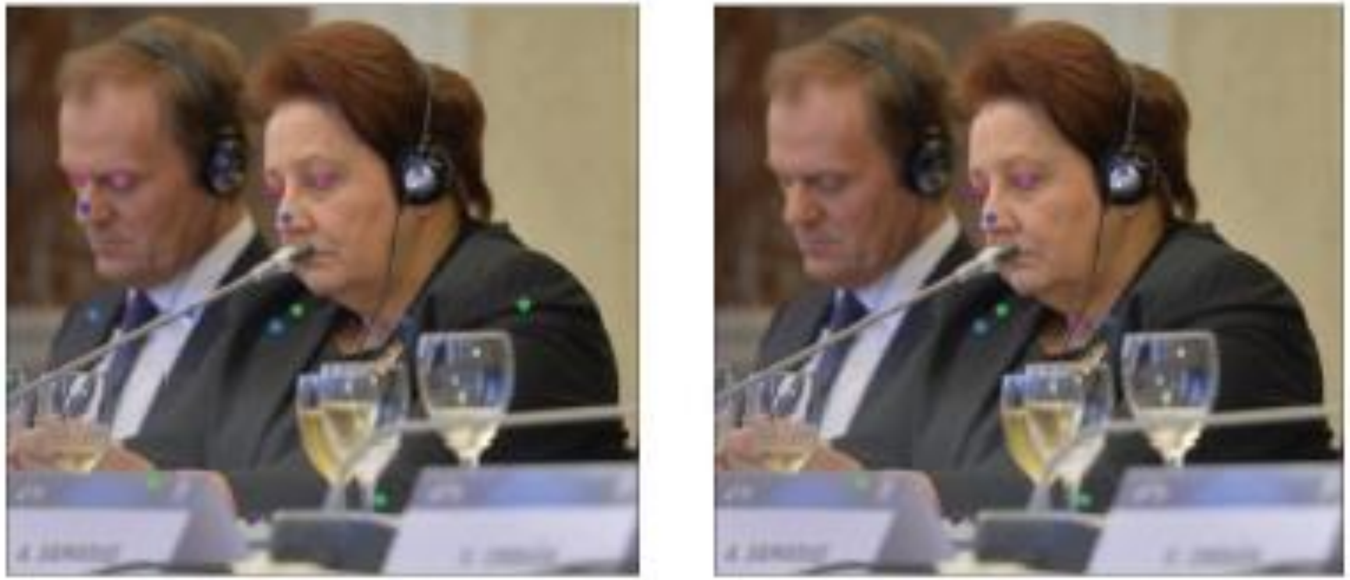
Baseline models

- Cropped image resized to 256x256 and fed into logistic regression model
- Second classifier is another logistic regression with # of people and size/location of each person

Pre-trained models

- ResNet50 – object detection and recognition model for 14M images
  - Replace final layer with fully connected sigmoid layer – only update parameters of new layer
- OpenPose – estimate body pose of person
  - Detect 18 regions/joints of human body
  - For duplicates (>1 person), pick part closest to center
- Emotion features (Hu and Ramanan)

Refining body joints



(a) The colored dots show the body joints of the two people originally detected.

(b) Result of removing duplicate body joints based on the distance from image center.

Fig. 2. Detecting and refining body joints.

Why were people subjects?

TABLE I
MOST FREQUENT REASONS FOUND IN THE PILOT STUDY FOR CLASSIFYING A PERSON AS A *Subject* AND HOW MANY TIMES EACH OF THEM WAS SELECTED IN THE MAIN STUDY.

#	Reason	Frequency
1	This photo is focused on this person.	5091
2	This photo is about what this person was doing.	4700
3	This is the only person in the photo.	2740
4	This person is taking a large space in the photo.	2425
5	This person was doing the same activity as other subject(s) in this photo.	2357
6	This person was interacting with other subject(s) in this photo.	1715
7	The appearance of this person is similar to other subject(s) of this photo.	1644

Why were people bystanders?

TABLE II
MOST FREQUENT REASONS FOUND IN THE PILOT STUDY FOR CLASSIFYING A PERSON AS A *Bystander* AND HOW MANY TIMES EACH OF THEM WAS SELECTED IN THE MAIN STUDY.

#	Reason	Frequency
1	This photo is not focused on this person.	3553
2	This person just happened to be there when the photo was taken.	2480
3	The activity of this person is similar to other bystander(s) in this photo.	1758
4	Object(s) other than people are the subject(s) of this photo.	1644
5	Appearance of this person is similar to other bystanders in this photo.	1278
6	There is no specific subject in this photo.	849
7	This person is interacting with other bystander(s).	755
8	This person is blocked by other people/object.	567
9	Appearance of this person is different that other subjects in this photo.	537
10	The activity of this person is different than other subjects(s) in this photo.	466

Second (test) dataset

- 600 images from Common Objects in Context (COCO)
- More mechanical Turk, but different participants

Predicting Survey Answers

- Predict Pose, Replaceable and Photographer’s intention
  - Use pre-trained models to guess these

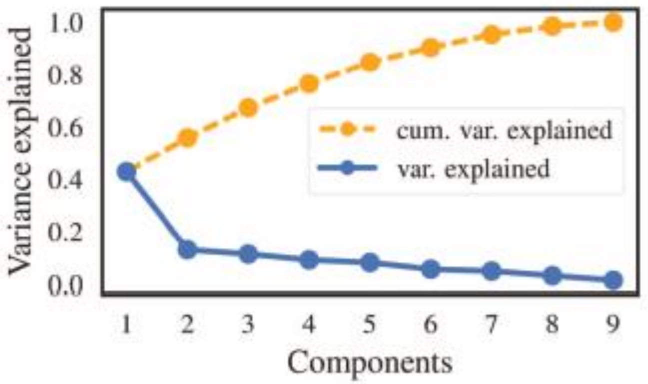
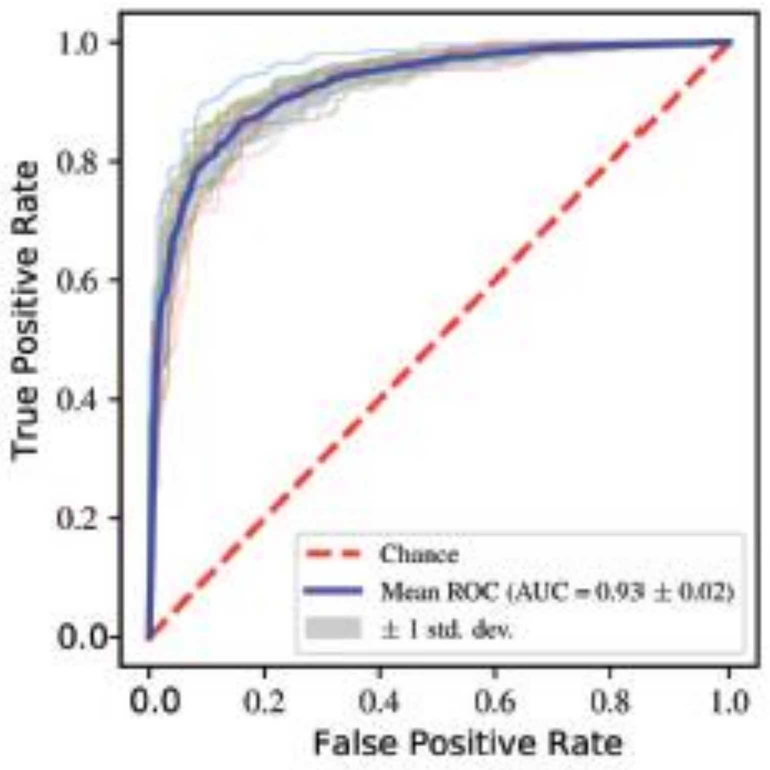


Fig. 3. Scree plot showing *proportions of variance* and *cumulative proportion of variance* explained by each component extracted using PCA.

Results (ROC)



ı-(f) Predicted *Pose*, *Replaceable*, *Photographer's intention*, and *Size*

Results (Accuracy)

TABLE VI
MEAN AND STANDARD DEVIATION OF ACCURACY FOR CLASSIFICATION USING DIFFERENT FEATURE SETS ACROSS 10-FOLD CROSS VALIDATION.

Features	Accuracy Mean	Accuracy SD
Cropped image	66%	0.03
Size, *distance*, and *number of people*	76%	0.01
Fine-tuning *ResNet*	77%	0.02
ResNet, *Pose*, and *Facial expression* features	78%	0.03
Size and ground truth *Pose*, *Replaceable*, *Photographer's intention*	86%	0.04
Size and predicted *Pose*, *Replaceable*, *Photographer's intention*	85%	0.02

More on results

- Accuracy was 93% when humans agree, but 80% when 2/3 humans agreed

To dos

- Cross-cultural analysis
- Use features from multiple people as predictors
- Use captions/friends list, etc.

COMPUTER SCIENCE & ENGINEERING
TEXAS A&M UNIVERSITY

Browsing Unicity: Limits of Anonymizing Web Tracking Data

Dr. Martin "Doc" Carlisle

Premise

- Anonymized browsing data can be de-anonymized

Cookies

- Third-party cookies allow publishers to track visits across websites
- Used for selling ads, e.g.

Privacy concerns

- Medical advice
- Planned parenthood
- Political discussion
- Pornographic content
- …

Threats to pseudonymity

- Tracking companies remove IP addresses, URL parameters, etc.
- But,
  - What if I can correlate with your visits to my site?
  - What if I shoulder surf you briefly
    - Possibly even using your public social media posts?
  - What if multiple tracking companies collaborate?

K-anonymity

- A database is 2-anonymous if no click trace is unique
  - Unlikely

Unicity

- Proportion of unique pieces of information
- 0 is k-anonymous, k>=2
- 0.25 means 1/4 of the click traces are unique

- Unique in the Crowd: The privacy bounds of human mobility (de Montjoye et al)
  - 4 spatio-temporal points uniquely identify 95% of individuals

Identifiability

- Chance you can obtain full trace from partial trace
- 0.2 means corresponding full trace has 20% chance to be identified

Definition 3 (identifiability): The *compatibility class* $\theta(\beta, T)$ of click trace β given traceset T consists of all click traces $\alpha \in T$ such that $\beta \subseteq \alpha$. We say that a click trace $\alpha \in T$ is *identified* by β, or β identifies α, if α is the only member of its compability class, or $\theta(\beta, T) = \alpha$. Given traceset I_α, the *identifiability* $\rho_\alpha(T, I_\alpha)$ of click trace $\alpha \in T$ is the ratio of click traces $\beta \in I_\alpha$ that α is identified by.

The weighted identifiability of a trace set T given $I = \{I_\alpha | \alpha \in T\}$ is

$$\rho(T, I) = \frac{\sum_{\alpha \in T}(|\alpha| \rho_\alpha(T, I_\alpha))}{\sum_{\beta \in T} |\beta|}$$

Creating Click Traces

- Push clicks from chronological click stream until two are more than 30 mins apart or exceeds max length

input : chronologically sorted stream C, max length ml;
all $c \in C$ contain timestamp c_t and click trace ID c_i
output: traceset T
$T \leftarrow \{\}$; TempTraces $\leftarrow \{\}$; LastTime $\leftarrow \{\}$;
for $c \in C$ **do**
  if $c_i \in \mathit{TempTraces}$ **and** $c_t - \mathit{LastTime}[c_i] < 1800$ **and** $\mathit{TempTraces}[c_i] < ml$ **then**
    TempTraces$[c_i] \leftarrow$ TempTraces$[c_i] \cup c$;
  else
    $T \leftarrow T \cup$ TempTraces$[c_i]$;
    TempTraces$[c_i] \leftarrow c$;
  end
  LastTime$[c_i] \leftarrow c_t$;
end
for $\mathit{trace} \in \mathit{TempTraces}$ **do**
  $T \leftarrow T \cup$ trace;
end

Algorithm 1: Calculating click traces from data stream

Calculating Unicity

- Use hashing set

input : traceset T, click trace properties w, hash function h
output: unicity and anonymity sets $Anon$ of T

```
Anon ← {}
for w_i ∈ w do
    for t ∈ T(w_i) do
        /* check if t's anonymity set already exists*/
        if t ∈ Anon then
            Anon(t) ← Anon(t) +1;
        else
            Anon(t) ← 1;
        end
    end
end
unique ← 0;
for t ∈ Anon do
    if Anon(t) = 1 then
        unique ← unique + 1;
    end
end
unicity ← unique / |T|
```

Algorithm 2: Unicity and anonymity sets given a traceset

Calculating identifiability

- Can’t use hashing trick as we have to determine if small set is part of larger one, or if equal
- Calculating for 3 observations on 1M traces of length 10 requires $14.4*10^{15}$ ops
- So we do sampling!

Bernoulli trials

- Pick random click (this picks a click trace weighted by its length)
- Select from all possible attacks

$$n_0 = \frac{Z^2 p(1-p)}{e^2}$$

- We don't know p, but p=0.5 maximizes n
- 99% (Z=2.576) chance of max error 1% (e) yields n=16590

Anonymization

- Truncate IP addresses
- Truncate timestamps
- Truncate URL

Dataset

- German websites (audience measurement)
- 2-3B page impressions per day
- One week from March 2019- desktop only

Field	Content
Timestamp	Unix timestamp in microseconds
Client ID	Unique per user / browser, from cookie
Site	ID of visited website/FQDN
Code	ID of displayed page, assigned by publisher
Category	Category of page, according to ABC
Geolocation	DB lookup of client IP

TABLE I
INFORMATION STORED PER CLIENT ACTION

Dataset

- Sampled 1/16th of clients randomly, half of available sites
  - Resource limitations (Hadoop platform with 2000 cores)
  - Ran experiments on increasing sizes and saw convergence

PIs	Visits	Clients	Locations	Sites	Codes	Categories
147.9M	22.1M	4.1M	3053	1281	62.5K	725

TABLE II
COMPOSITION OF THE TESTED SAMPLE

Click trace unicity vs coarsened time

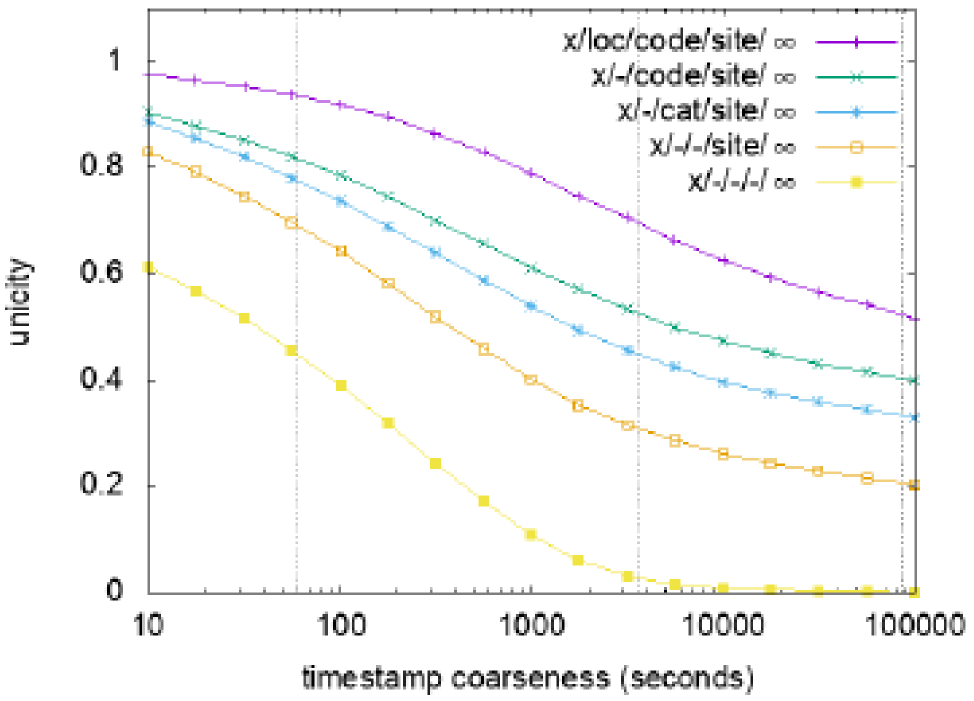


Fig. 2. Click trace unicity over coarsened time.

Unicity vs trace length

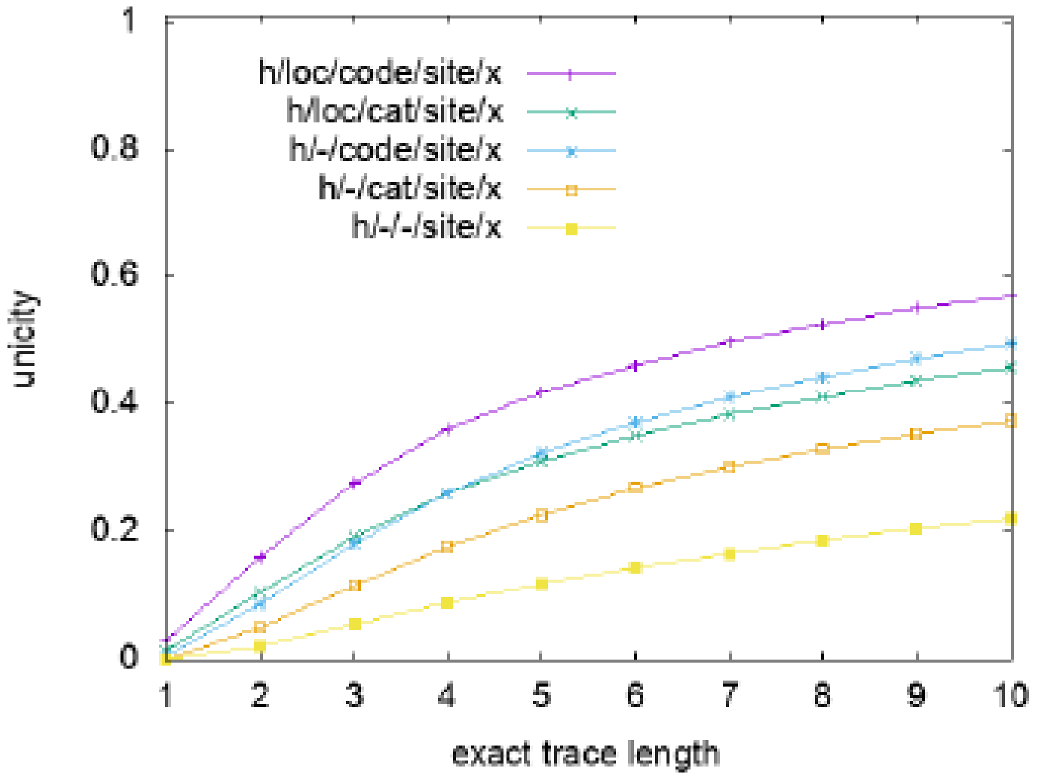


Fig. 5. Click trace unicity for exact trace length, timestamps coarsened to the hour.

Identifiability given known clicks

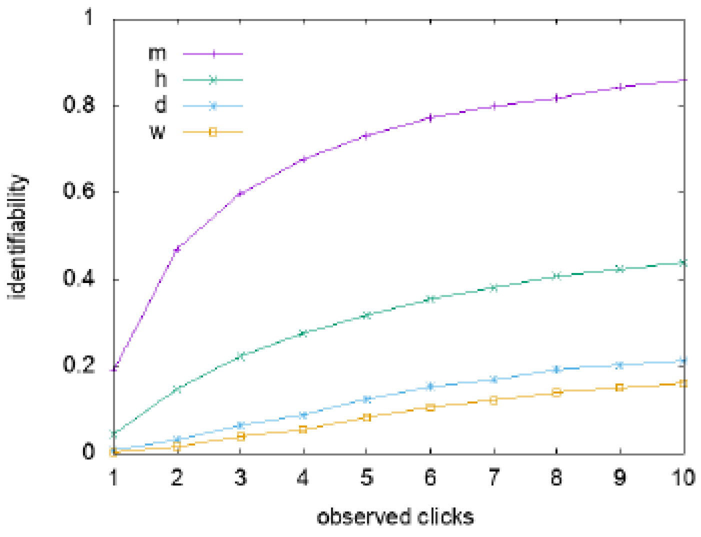


Fig. 10. Shoulder surfing: We measure the identifiability of a partially observed browsing session, given the number of observations. Configuration: -/loc/-/site/10.

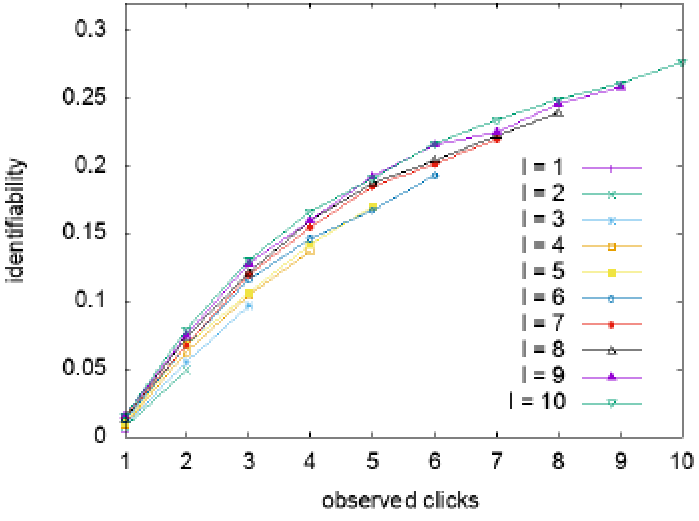


Fig. 11. Shoulder surfing: We measure the identifiability of a partially observed browsing session, given the number of observations for different session lengths. Configuration: h/loc/-/site/-.

How to get < 10% unicity

- Remove all info pertaining to clients and website visits
- Coarsen time to at least hours