SocialSpamGuard

Dr. Martin “Doc” Carlisle
• 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

![Diagram of 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
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

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".

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.
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 number 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 1

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.

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

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

- 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.

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cropped image</td>
<td>66% 0.03</td>
</tr>
<tr>
<td>Size, distance, and number of people</td>
<td>76% 0.01</td>
</tr>
<tr>
<td>Fine-tuning ResNet</td>
<td>77% 0.02</td>
</tr>
<tr>
<td>ResNet, Pose, and Facial expression features</td>
<td>78% 0.03</td>
</tr>
<tr>
<td>Size and ground truth Pose, Replaceable, Photographer’s intention</td>
<td>86% 0.04</td>
</tr>
<tr>
<td>Size and predicted Pose, Replaceable, Photographer’s intention</td>
<td>85% 0.02</td>
</tr>
</tbody>
</table>
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.
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 \( \beta \) 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 \( \beta \), or \( \beta \) identifies \( \alpha \), if \( \alpha \) is the only member of its compatibility class, or \( \theta(\beta, T) = \alpha \). Given traceset \( I_\alpha \), the **identifiability** \( \rho_\alpha(T, I_\alpha) \) of click trace \( \alpha \in T \) is the ratio of click traces \( \beta \in I_\alpha \) that \( \alpha \) 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 \( c_i \in TempTraces \) and \( c_i - LastTime[c_i] < 1800 \) and
        \( TempTraces[c_i] < ml \) 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 trace \( \in TempTraces \) do
    \( T \leftarrow T \cup trace; \)
end
Algorithm 1: Calculating click traces from data stream
```
Calculating Unicity

• Use hashing set

```
input : trace set T, click trace properties w, hash function h
output: unicity and anonymity sets Anon of T
Anon ← {}
for \( w_i \in w \) do
    for \( t \in T(w_i) \) do
        /* check if t’s anonymity set already exists*/
        if \( t \in \text{Anon} \) then
            Anon(t) ← Anon(t) + 1;
        else
            Anon(t) ← 1;
        end
    end
end
unique ← 0;
for \( t \in \text{Anon} \) do
    if \( \text{Anon}(t) = 1 \) then
        unique ← unique + 1;
    end
end
unicity ← \[ \frac{\text{unique}}{|\text{Anon}|} \]
```

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 $14.4 \times 10^{15}$ 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

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

**Table I**

Information stored per client action
Dataset

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

<table>
<thead>
<tr>
<th>PIs</th>
<th>Visits</th>
<th>Clients</th>
<th>Locations</th>
<th>Sites</th>
<th>Codes</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>147.9M</td>
<td>22.1M</td>
<td>4.1M</td>
<td>3053</td>
<td>1281</td>
<td>62.5K</td>
<td>725</td>
</tr>
</tbody>
</table>

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: /loc/-/site/10.
How to get < 10% unicity

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