Sharing Multimedia on the Internet and the Impact for Online Privacy

Dr. Gerald Friedland
Senior Research Scientist
International Computer Science Institute
Berkeley, CA
friedland@icsi.berkeley.edu
Question

On average, how often are you posting images and videos on the Internet (e.g. Facebook, Flickr, Craigslist)?

a) Never
b) About once a month or less
c) About once a week
d) About once a day
e) More than once a day
A Popular Introduction to the Problem
How would you judge the issue raised by Colbert?

a) It’s a comedy. I don’t worry about any of this.

b) There is some truth to it but its mostly exaggerated.

c) It’s a comedy depection of the reality but most of the stuff is becoming an issue.

d) He only touched a small part of the problem. The actual issues are even more serious.
Our Observations

• Many Internet sites and mobile apps encourage sharing of data too easily and users follow.
• Users and even engineers often unaware of (hidden) search and retrieval possibilities of shared data.
• Local privacy protection ineffective against inference across web-sites.
Social Cause

• People want to post on the Internet and like a highly-personalized web experience.
• Industry is improving search and retrieval techniques so that people can find the posts.
• Governments improve search and retrieval to do forensics and intelligence gathering
Let’s focus

• The previous described issues are a problem with any type of public or semi-public posts and are not specific to a certain type of information, e.g. text, image, or video.

• However, let’s focus on multimedia data: images, audio, video.
Multimedia in the Internet is Growing

- YouTube claims 65k-100k video uploads per day, or 48h-72h per minute.

- Flickr claims 1M images uploads per day

- Twitter: up to 120M messages per day
• More multimedia data = Higher demand for retrieval and organization tools.

• But multimedia retrieval is hard
  ➔ Researchers work on making retrieval better.
  ➔ Industry develops workarounds to make retrieval easier right away.
• Retrieval is already good enough to cause major issues for privacy that are not easy to solve.

• Let’s take a look at some retrieval approaches:
  - Geotagging
  - Multimodal Location Estimation
  - Audio–based user matching
Workaround: Manual Tagging
Workaround: Geotagging

Geo-Tagging

Allows easier clustering of photo and video series as well as additional services.
Support for Geo−Tags

Social media portals provide APIs to connect geo−tags with metadata, accounts, and web content.

<table>
<thead>
<tr>
<th>Portal</th>
<th>%</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>YouTube</td>
<td>3.0</td>
<td>3M</td>
</tr>
<tr>
<td>(estimate)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flickr</td>
<td>4.5</td>
<td>180M</td>
</tr>
</tbody>
</table>

Allows easy search, retrieval, and ad placement.
“Be careful when using social location sharing services, such as FourSquare.”
Related Work

Mayhemic Labs, June 2010: “Are you aware that Tweets are geo-tagged?”
Question

Did you know about geo-tagging and it’s potential?

a) I had never heard about geo-tagging before.
b) I knew about geo-tagging but never thought about what it could be used for.
c) I knew about geo-tagging and knew the potential for photo organization and retrieval
d) I know about geo-tagging, it’s use and the privacy risks.
e) I only heard about privacy risks of geo-tagging but never really thought about what it is good for.
Since geo-tagging is a workaround for multimedia retrieval, it allows us to peek into a future where multimedia retrieval works.

What if multimedia retrieval actually worked?
Can you do real harm?

- **Cybercasing:** Using online (location-based) data and services to enable real-world attacks.

- **Three Case Studies:**

Case Study 1: Twitter

- Pictures in Tweets can be geo-located
- From a technically-savvy celebrity we found:
  - Home location (several pics)
  - Where the kids go to school
  - The place where he/she walks the dog
  - “Secret” office
Celebs unaware of Geo-Tagging

Working with the very talented Adam Hamilton on creating a new album. My best, Bill

Source: ABC News
Celebs unaware of Geotagging

**EXIF IFD1**

- Compression {0x0103} = JPEG compression (6)
- X-Resolution {0x011A} = 4718592/65536 ===> 72
- Y-Resolution {0x011B} = 4718592/65536 ===> 72
- X/Y-Resolution Unit {0x0128} = inch (2)
- Y/Cb/Cr Positioning (Subsampling) {0x0213} = centered / center of pixel array (1)
- Embedded thumbnail image:

**EXIF GPS IFD**

- GPS Version ID {0x00} = 0x02,0x02,0x00,0x00
- GPS Latitude Reference {0x01} = N
- GPS Latitude {0x02} = 34/1,12/1,3/1 [degrees, minutes, seconds] ===> 34° 12’ 3” = 34.200833°
- GPS Longitude Reference {0x03} = W
- GPS Longitude {0x04} = [degrees, minutes, seconds] ===> -118°? “ = -118°?
Google Maps shows Address...
Case Study 2: Craigslist

“For Sale” section of Bay Area Craigslist.com: 4 days: 68729 pictures total, 1.3% geo-tagged

<table>
<thead>
<tr>
<th>#</th>
<th>Model</th>
<th>#</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>414</td>
<td>iPhone 3G</td>
<td>6</td>
<td>Canon PowerShot SD780</td>
</tr>
<tr>
<td>287</td>
<td>iPhone 3GS</td>
<td>3</td>
<td>MB200</td>
</tr>
<tr>
<td>98</td>
<td>iPhone</td>
<td>2</td>
<td>LG LOTUS</td>
</tr>
<tr>
<td>32</td>
<td>Droid</td>
<td>2</td>
<td>HERO200</td>
</tr>
<tr>
<td>26</td>
<td>SGH-T929</td>
<td>2</td>
<td>BlackBerry 9530</td>
</tr>
<tr>
<td>20</td>
<td>Nexus One</td>
<td>1</td>
<td>RAPH800</td>
</tr>
<tr>
<td>9</td>
<td>SPH-M900</td>
<td>1</td>
<td>N96</td>
</tr>
<tr>
<td>9</td>
<td>RDC-i700</td>
<td>1</td>
<td>DMC-ZS7</td>
</tr>
<tr>
<td>6</td>
<td>T-Mobile G1</td>
<td>1</td>
<td>BlackBerry 9630</td>
</tr>
</tbody>
</table>
People are Unaware of Geo-Tagging

- Many ads with geo-location otherwise anonymized
- Sometimes selling high-valued goods, e.g. cars, diamonds
- Sometimes “call Sunday after 6pm”
- Multiple photos allow interpolation of coordinates for higher accuracy
Craigslist: Real Example
Case Study 3: YouTube

Recall:
- Once data is published, the Internet keeps it (in potentially many copies).
- APIs are easy to use and allow quick retrieval of large amounts of data.

Can we find people on vacation in YouTube?
Cybercasing on YouTube
Experiment: Cybercasing using the YouTube API (240 lines in Python)
Cybercasing on YouTube

Input parameters

Location: 37.869885, -122.270539
Radius: 100km
Keywords: kids
Distance: 1000km
Time-frame: this_week
Cybercasing on YouTube

First Day of Vacation

videos Subscribe
The Threat is Real!

Bits

Business • Innovation • Technology • Society

September 12, 2010, 10:24 AM

Burglars Picked Houses Based on Facebook Updates

By NICK BILTON

Illustration by Nick Bilton/The New York Times
Do you think geo-tagging should be illegal?

a) No, people just have to be more careful. The possibilities still outweigh the risks.

b) Maybe it should be regulated somehow to make sure no harm can be done.

c) Yes, absolutely this information is too dangerous.
But...

Is this really about geo-tags? (remember: hypothesis)
Ongoing Work:

The Berkeley Multimodal Location Estimation Project

http://mmle.icsi.berkeley.edu
Multimodal Location Estimation

We infer location of a Video based on visual, audio, and tags:

- Use geo-tagged data as training data
- Allows faster search, inference, and intelligence gathering even without GPS.

## YouTube Cybercasing Revisited

<table>
<thead>
<tr>
<th></th>
<th>Old Experiment</th>
<th>No Geotags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Videos</td>
<td>1000 (max)</td>
<td>107</td>
</tr>
<tr>
<td>User Hull</td>
<td>~50k</td>
<td>~2000</td>
</tr>
<tr>
<td>Potential Hits</td>
<td>106</td>
<td>112</td>
</tr>
<tr>
<td>Actual Targets</td>
<td>&gt;12</td>
<td>&gt;12</td>
</tr>
</tbody>
</table>

YouTube Cybercasing with Geo-Tags vs Multimodal Location Estimation
Question

Do you think research about geo-location should be abandoned?

a) No, of course not.
b) No, but regulated.
c) Yes, absolutely.
But...

Is this really only about geo–location?

No, it’s about the privacy implications of multimedia retrieval in general.
Example

Idea: Can one link videos across accounts? (e.g. YouTube linked to Facebook vs anonymized dating site)
User ID on Flickr videos

User-ID, i-vector system, 6.4 million trials (3287 true positive trials), 1239 train users, 2784 test users, 60–16000 Hz MFCC

- 68.1% miss at 1% false alarm
- 26.3% EER
Result:
On average having 20 videos in the test set leads to a 99% chance for a true positive match!

And now? What do you think has to be done?

a) Nothing can be done. Privacy is dead.
b) We need to educate people about this and try to save privacy. (fight)
c) I will really think before I post, and I agree with b).
d) I will really think before I post, and I agree with a).
e) I won’t post anything anymore! (flee)
More examples and more discussion

http://cybercasing.blogspot.com
Thank You!

Questions?

Work together with: