Exploiting Innocuous Activity for Correlating Users Across Sites

Oana Goga
UPMC Sorbonne Universités
23 Avenue d'Italie
Paris, France
oana.goga@lip6.fr

Howard Lei
ICSI
1947 Center St., Suite 600
Berkeley, USA
hlei@icsi.berkeley.edu

Sree Hari Krishnan Parthasarathi
ICSI
1947 Center St., Suite 600
Berkeley, USA
sparta@icsi.berkeley.edu

Gerald Friedland
ICSI and UC Berkeley
1947 Center St., Suite 600
Berkeley, USA
fractor@icsi.berkeley.edu

Robin Sommer
ICSI and LBNL
1947 Center St., Suite 600
Berkeley, USA
robin@icir.org

Renata Teixeira
CNRS and UPMC Sorbonne Universités
23 Avenue d'Italie
Paris, France
renata.teixeira@lip6.fr

WWW 2013

Tehila Minkus

Lots of public data in lots of places

More than the sum of their parts?

Attempting account linkage across

flickr

yelp

The tools:

YES:

NO:

username

Threat model

- Attacker with moderate resources
- Given account a in SN1, wants to find corresponding account a in SN2
- Attacker can’t crawl entire network so needs to limit himself to a subgroup of SN2

Attack algorithm

1. Select a target from OSN 1
2. Filter entire search space of OSN 2
3. Generate similarity score for each candidate
4. Output top (or top k) matches
5. If this is (includes) the match, success!

Collecting ground truth

- 10,000 email addresses from previous study
- Use browser automation to find corresponding accounts with “FriendFinder” feature
- Limit dataset to accounts with geotagged data

Collecting potential matches

- Twitter: use Streaming API to collect tweets in specific geographic areas
- Flickr: collect photos with geotags in those areas
- Yelp: collect reviews from restaurants in those areas

Geo Subgroups for Matching

	GT	GT in SF†	SD†	NY†	C†	LA†
Twitter-Flickr	13,629	474	152	427	236	284
Twitter-Yelp	1,889	160	45	106	50	117
Flickr-Yelp	1,199	120	46	81	42	82
Twitter-Flickr-Yelp	559	33	9	25	11	23

Table 1: Number of users in the ground-truth dataset GT (total, and divided into 5 selected areas). † Users with more posts inside a given area than outside it.

	$\widetilde{\text{SN}}_2$	$\widetilde{\text{SN}}_2$ in SF†	SD†	NY†	C†	LA†
Twitter	232,924	75,747	35,068	89,219	54,774	77,402
Flickr	22,169	6,916	2,305	5,730	4,122	4,113
Yelp	28,976	16,463	4,064	6,239	3,629	9,556

Table 2: Number of users in the $\widetilde{SN}_2$ dataset (total, and divided into 5 selected areas). † Users with at least one post inside a given area; users may belong to multiple areas.

Features

- Location profile: histogram of clustered places from which a user has posted, normalized to represent prob. distribution of locations
- Time profile: windows of times when user has posted
- Language profile: a prob. distribution based on unigram histogram

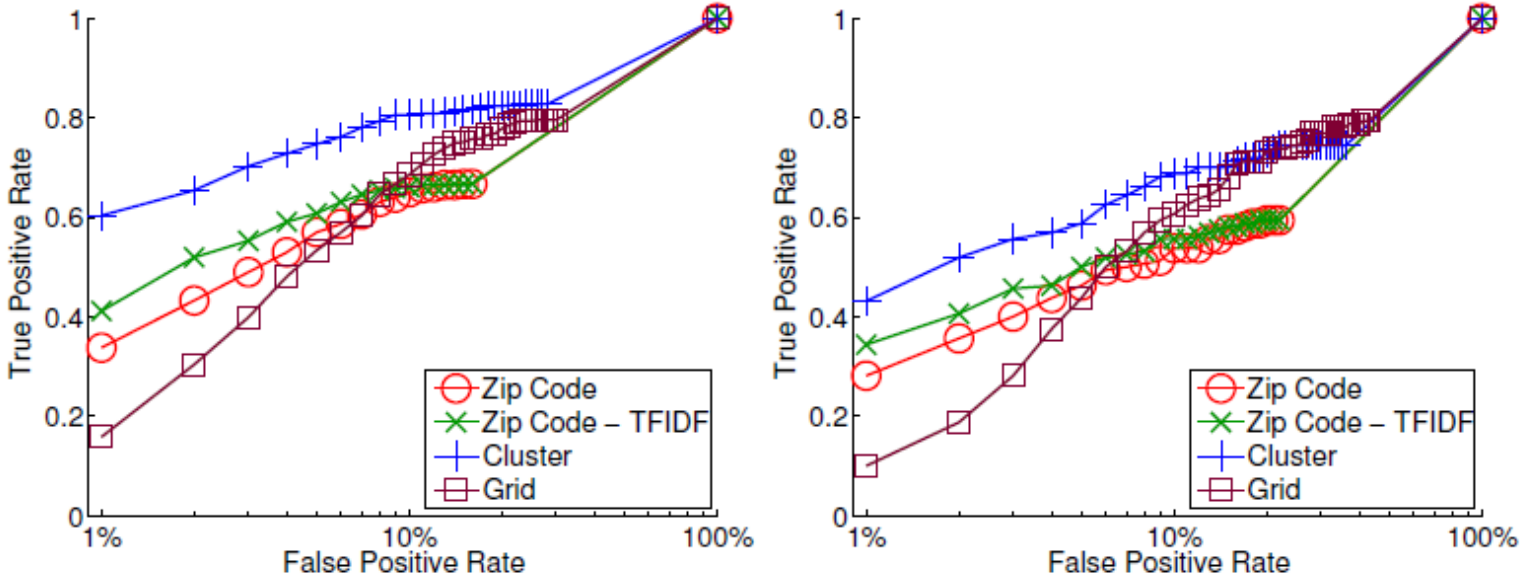
Fixing Location: Approaches

- Grids (10x10 km)
- Zip code
- Zip codes weighted by TF-IDF
- Clustered locations (unsupervised)

Location clustering, detail

- Run k-means clustering on all the geotagged data in each city to find 10,000 *landmark clusters*
- Then represent a user's geotag data as a weighted distribution of its 20 closest landmarks
- Each user's location profile is a histogram based on all of his geotags

Location Approaches: Comparison



(a) Flickr to Twitter

(b) Yelp to Twitter

Figure 1: ROC curves for different location representations for matching Flickr and Yelp users (GT_1^{SF}) to Twitter users ($\widetilde{SN}_2^{SF}$).

Approaches to Time and Language

- Time: a sliding window of 5 seconds to match posts (looking for automatic posts)
- Language: remove case sensitivity and punctuation, remove 1000 top words, then consider unigrams (better performance than n-grams)

Combining Features

Uses binary logistic regression classifier that takes similarity score and outputs “match/no match” and probability of matching

Results

Table 3: Comparison of the TPR for different classifiers at 1% FPR for matching Flickr and Yelp accounts to Twitter.

Feature	TPR at 1%FPR	
	Flickr-Twitter	Yelp-Twitter
Timing (T)	13±3%	-
Language (Lang)	10±3%	6±3%
Location (Loc)	60±6%	44±6%
Username (U)	77±3%	7±4%
Loc, Lang	60±6%	42±6%
Loc, T	70±3%	-
Loc, Lang, T	63±5%	-
Loc, U	86±2%	44±6%
Loc, Lang, U	86±2%	44±7%
Loc, T, Lang, U	88±2%	-

Discussion and Future Work

- Extending to other OSNs
- Considering attacks against groups rather than individuals
- Inferring location data from other posts, etc
- How to mitigate risk?
  - no automatic posting
  - don’t post to several sites from same location

Thanks!

Full paper available at:

http://www.icir.org/robin/papers/www13-correlation.pdf