Spot the Hotspot:
Wi-Fi Hotspot Classification from Internet Traffic

Andrey Finkelshtein
Rami Puzis
Asaf Shabtai
Bronislav Sidik

Ben Gurion University of the Negev

SMARTPHONES

- Smartphones are everywhere
  - Smartphones are very common
  - The smartphone is always with us
- Not only we need the smartphone –
  We need to be online all the time

SMARTPHONE USAGE

Usage of mobile phones by mobile phone users in the UK (2013) [2]

- Phone calls and Texting (98%)
- Take photos (67%)
- Browse websites (54%)
- E-mails (51%)
- Social Networks (51%)
- Download apps and games (39%)
- Listen to Music (41%)
- …

Requires Internet

3

[2] *Adults' Media Use and Attitudes Report.* Ofcom (2014)

SMARTPHONE USAGE

- People use their smartphones for different purposes in different venues
  - Will you stream music during a lecture?
  - Will you play mobile games while sitting in a bar with friends?
  - Will you video chat with a colleague while taking a bus?

The venue of a smartphone user can be defined as the context of a user

4

VENUE AS A CONTEXT

- Context based recommendation systems:
  - Recommend downloading a game to a user waiting in line
- Context based preferences:
  - Larger fonts for a bus passenger
  - Long articles for a user at home
- Context based access control:
  - Do not expose sensitive information to a user in a public place
- Context based commercials:
  - A beverage to a user dinning in a restaurant

5

HOW TO FIND THE VENUE?

- Ask the user
- Derive it
  - **User's location**
    - Not always accurate or available
    - Requires knowledge about the physical location
    - Different venues can be located next each or even in the same building
  - **Smartphone sensors (Wi-Fi scanner, light sensor, accelerometer…)**
    - Requires presence on the smartphone (Application)
    - Permissions, battery and performance issues
  - **Internet Traffic**
    - Does not require presence on the smartphone
    - Does not bother the user

6

PROBLEM DEFINITION

- Given Internet traffic records of a smartphone user, find the type* of the venue the user is located in.

*We'll define venues types shortly

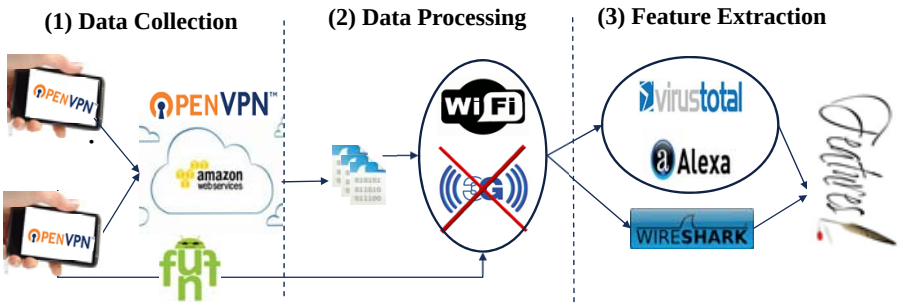
7

DATA COLLECTION

- Experiments with smartphone users:
  - Conducted during 2014 and 2015
  - A total of 115 student subjects
  - Information collected:
    - Internet traffic records from subjects' devices
    - Wi-Fi hotspots connection and disconnections information
  - Record periods of 30 to 60 consecutive days per user

8

DATA COLLECTION PROCESS



- (1) Client application redirects Internet traffic to a server and records Wi-Fi connection/disconnection events
- (2) Cellular traffic is filtered out
- (3) Features are extracted for each session

9

FROM WI-FI HOTSPOTS TO VENUES TYPES

- Wi-Fi hotspots are now everywhere
  - In our homes
  - In our neighbors houses
  - Restaurants
  - Shopping malls
  - Workplaces
  - Education institutes
  - Public transport
  - Hair salons

Manual labeling by SSID

- Home
- Hangout places
- Organizations
- Waiting places

10

INTERIM SUMMARY

- So far:
  - We saw venues are part of the context of smartphone users
  - We collected smartphones traffic records within hotspots
  - We labeled the hotspots by the type of the venues they are located in
    - Home, Organization, Hangout and Waiting
- What is next:
  - Extract features from traffic
  - Analyze the data
  - Build classification model

11

FEATURE EXTRACTION

- Communication Features
  - Statistics on traffic volumes, arrival times, lost packets, etc…
- Domains Info
  - Categories
  - Security scores
  - Popularity
- Protocols based Features
  - HTTP cookies
  - SSL certificates

12

DATA ANALYSIS

We distinguish between two different types of attributes:

- User behavior attributes:
  - Domains categories and popularity
  - Traffic volumes
  - …
- Hotspot properties attributes:
  - Packets arrival time
  - Packet loss
  - …

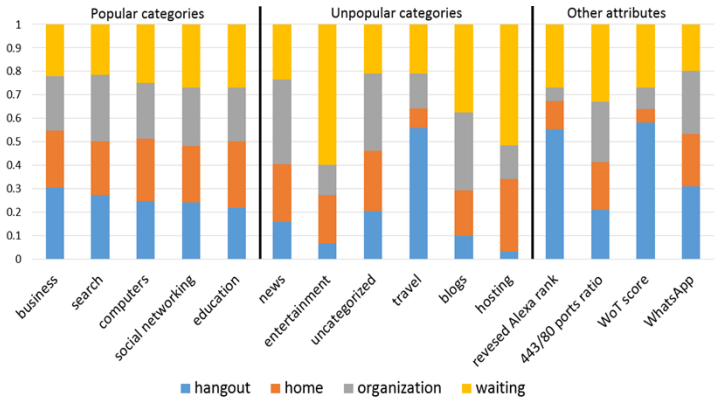
13

USER BEHAVIOR ANALYSIS

Proportional amount of traffic volumes of different behavioral attributes for each venue type:

No big difference in popular categories

Significant difference in the unpopular categories



Popularity scores and security scores seems to correlate

We hypothesize that a large amount of the traffic is generated by the device and not the user

14

HOTSPOT VENUES CLASSIFICATION

- Sessions are too small units to be classified
  - ➔ need to aggregate the data into larger chunks
- We aggregated the data into connection windows (CWs):
  - CW is defined as all of the session of a user within a single hotspot with an idle time smaller than 30 minutes

15

CONNECTION WINDOWS

- The aggregation resulted 37,714 CWs
- the majority of them are "Home" CWs (72.56%)
- Only 720 "Waiting" CWs (1.91%)
- To overcome the imbalance problem we created 3 dataset by sampling randomly CWs:
  - 3600 CWs in each dataset:
    - 960 Home CWs
    - 960 Organization CWs
    - 960 Hangout CWs
    - 720 Waiting CWs

Hotspot Type	Number of Connection Windows
Home (H)	27,367
Organization (O)	7,929
Hangout (HO)	2,708
Waiting (W)	720
Total	**37,714**

16

MULTICLASS MODEL

- A multiclass classification approach:
  - Feature selection – Correlation Feature Selection (CFS)
  - Classification Model – Rotation Forest Ensemble algorithm
  - Evaluation – 10-fold cross validation
  - Results:

	Dataset1	Dataset2	Dataset3
Accuracy	57.75%	57.67%	58.5%
Weighted AUC	0.81	0.82	0.83
Home AUC	0.75	0.73	0.76
Hangout AUC	0.90	0.90	0.90
Org. AUC	0.80	0.81	0.83
Waiting AUC	0.80	0.81	0.82

17

ONE VERSUS ALL MODEL

- Train 4 classifiers:
  - one for each hotspots type (venue)
  - The rest of the types are labeled as "other"
- The classification is made using the probabilities of the classifiers
  - Each classifier provides the probability of a CW to be from a specific type
  - The type is decided by the highest probability
- This model showed some improvement compared to the multiclass model
- Performed better in classifying "Home" and "Org." hotspots

18

COMBINED MODEL

- Use both models to classify CWs:
- Classify using the "1 vs. all" model
  - If the classification is "Home" or "Organization" use the classification
  - else, use the multiclass model to classify the CW
- Evaluation:
  - Evaluate on the same 3 datasets
  - Random train-test splits:
    - Test groups of 100, 200, 400 and 800 instances

19

COMBINED MODEL EVALUATION RESULTS

	Dataset1				Dataset2				Dataset3			
	H.	**O.**	**HO**	**W.**	**H.**	**O.**	**HO**	**W.**	**H.**	**O.**	**HO**	**W.**
Avg. Precision	.73	.55	.55	.63	.70	.50	.46	.67	.66	.62	.51	.70
Avg. Recall	.81	.66	.48	.50	.84	.61	.34	.57	.81	.71	.46	.51
Avg. F-measure	.77	.60	.51	.56	.76	.55	.39	.62	.73	.66	.48	.59
Accuracy	**76.5%**				**75.1%**				**78.1%**			

20

CONCLUSIONS AND FUTURE WORK

- Conclusions
  - Smartphone users' behavior changes across different types of venues
  - Internet traffic can be used to classify the type of the venue
  - The differences in the behavior are reflected in the less popular domain categories
- Future work
  - Improve the dataset
    - More subjects, Higher diversity
  - Classify venues using other data sources:
    - Use smartphone sensors available to mobile applications

21

QUESTIONS?

Thank you!

22