Practical Data Science – COSC2670

Practical Data Science:

Identifying In-App User Actions from Mobile Web Logs

Dr. Yongli Ren (with my colleagues)

(yongli.ren@rmit.edu.au)

Computer Science & IT

School of Science

RMIT UNIVERSITY

Outline

- Project
- Data
- Exploration
- Modelling
- Conclusion

Project

- We address the problem of
  - identifying in-app user actions from Web access logs
    - when
      - the content of those logs is both *encrypted* (through HTTPS) and
      - also contains *automated Web accesses*.
- We find that
  - the *distribution of time gaps* between HTTPS accesses can
    - *Distinguish* user actions from automated Web accesses
      - which generated by the apps.
- We determine that
  - it is reasonable to identify meaningful user actions within mobile Web logs
    - by modelling this temporal feature with DBSCAN.

Data

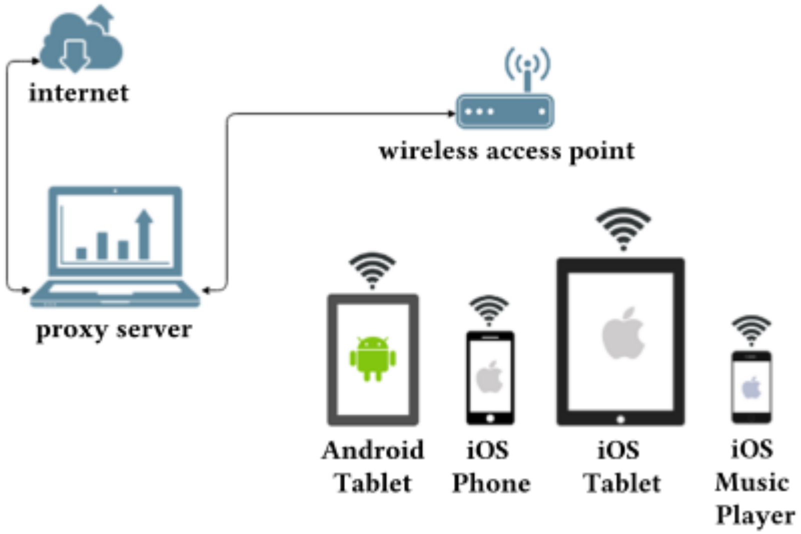
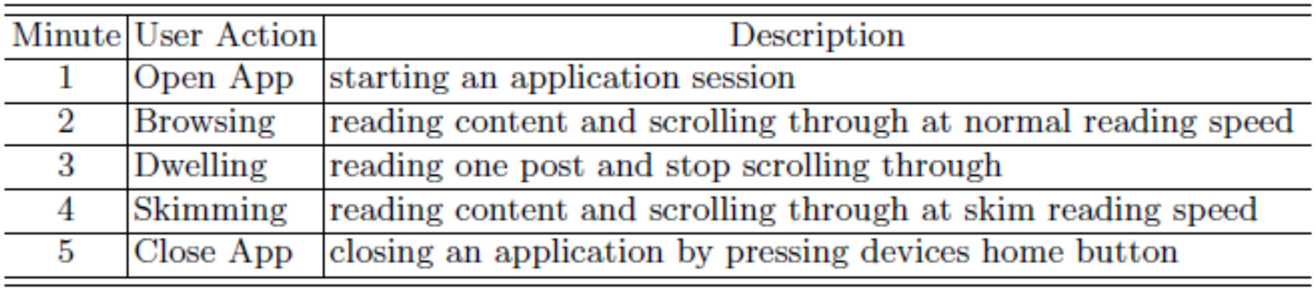
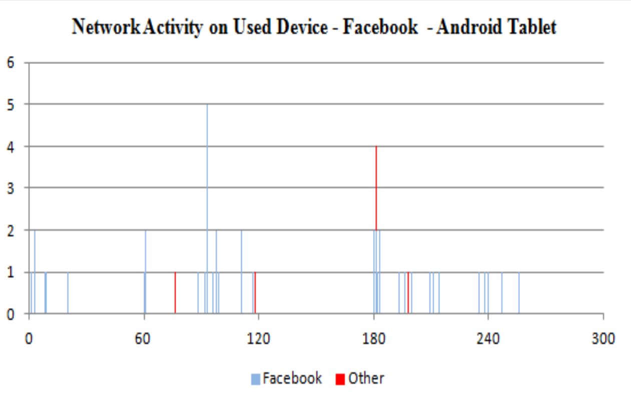
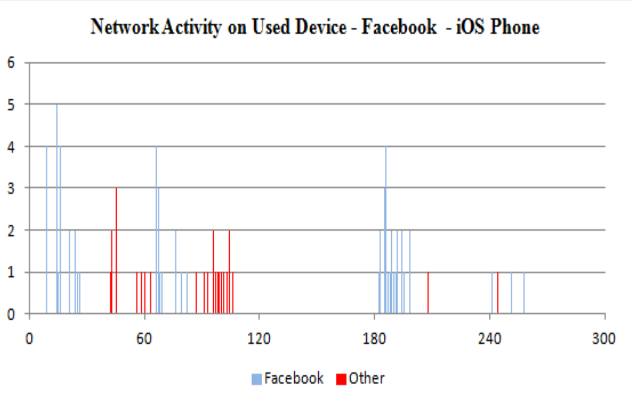


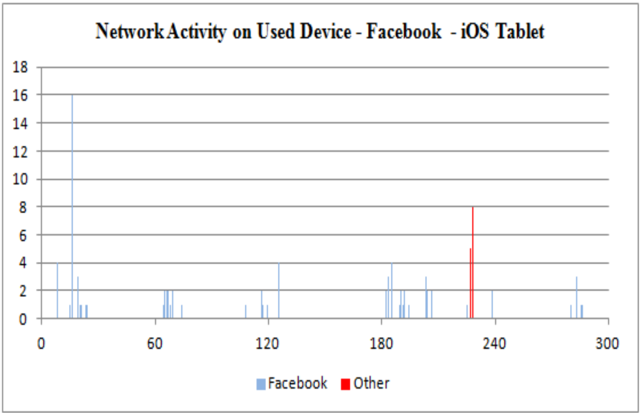
Table 2. User Actions Examined

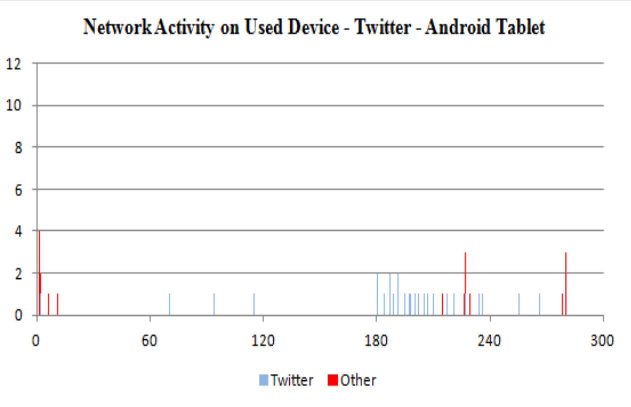


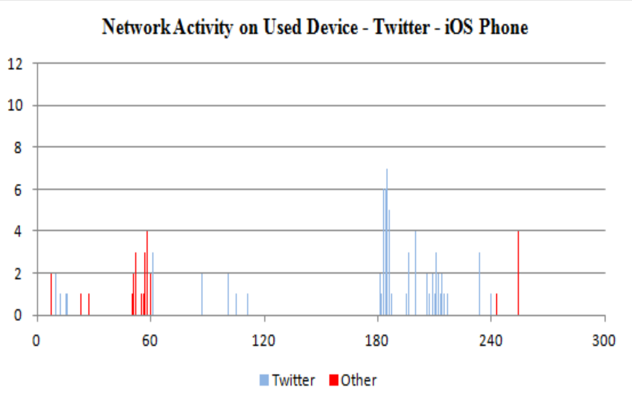
Minute	User Action	Description
1	Open App	starting an application session
2	Browsing	reading content and scrolling through at normal reading speed
3	Dwelling	reading one post and stop scrolling through
4	Skimming	reading content and scrolling through at skim reading speed
5	Close App	closing an application by pressing devices home button

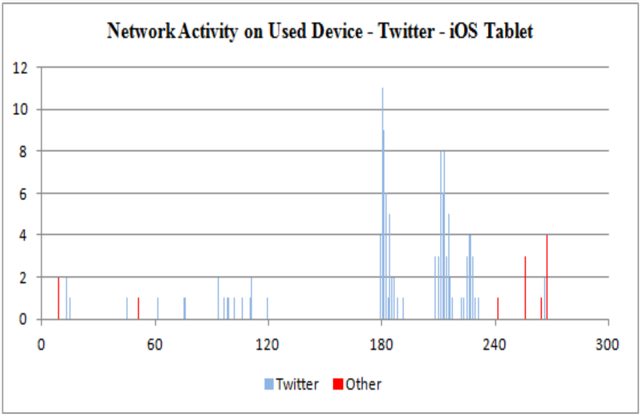

Network Activity on Used Device - Facebook - Android Tablet

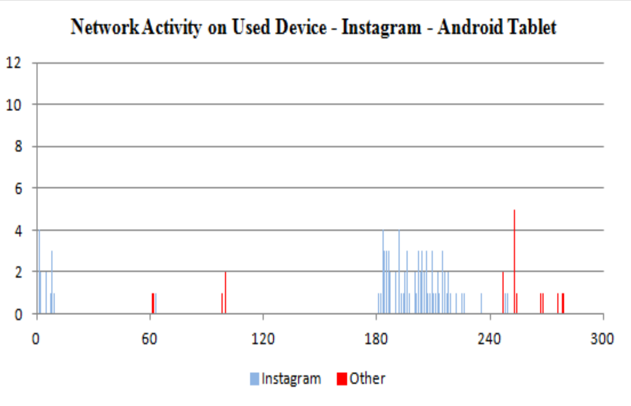

Network Activity on Used Device - Facebook - iOS Phone

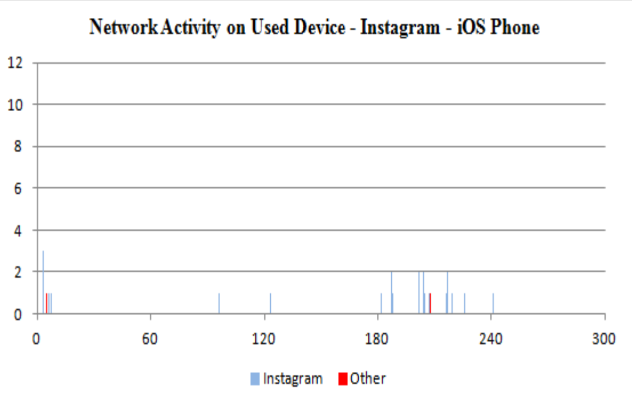

Network Activity on Used Device - Facebook - iOS Tablet

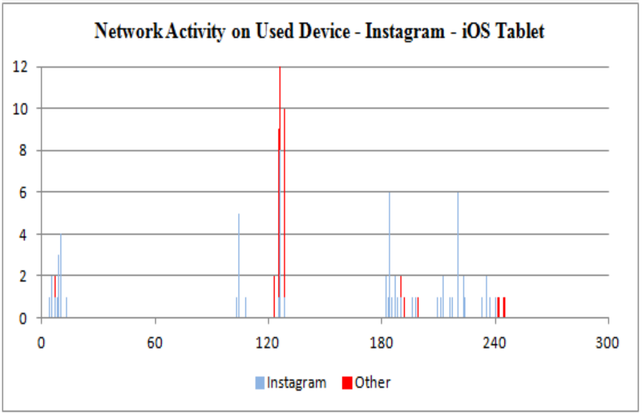

Network Activity on Used Device - Twitter - Android Tablet


Network Activity on Used Device - Twitter - iOS Phone


Network Activity on Used Device - Twitter - iOS Tablet


Network Activity on Used Device - Instagram - Android Tablet


Network Activity on Used Device - Instagram - iOS Phone

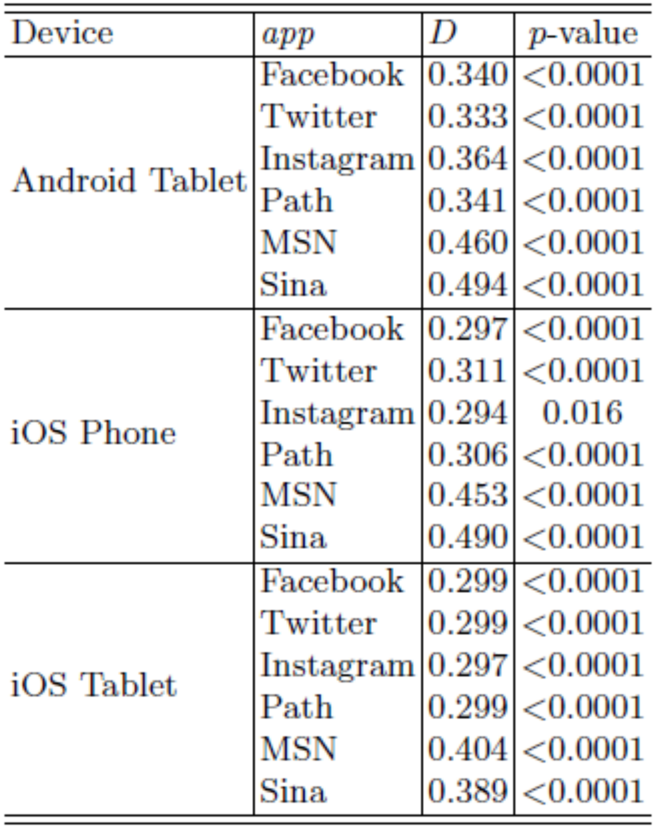

Network Activity on Used Device - Instagram - iOS Tablet

Exploration

- We first conducted
  - a *comprehensive analysis (exploration)* of *the time gap*,
    - which is defined as the gap in seconds
      - between *consecutive URL requests* from the *same device* and *app*.
- We examine the logs of six representative apps:
  - Facebook, Twitter, Instagram, Path, MSN, Sina;
  - on four different devices (Android Tablet, iOS Phone/Tablet/music player)
- The gaps are separated into two groups:
  - *idle* that means there are no user actions, and
  - *active* that means there are user actions with the device.

Exploration

- Statistical (Kolmogorov-Smirnov (KS)) tests are deployed to examine whether
  - the *idle vs active* time gap distributions are significantly.



Device	*app*	D	p-value
Android Tablet	Facebook	0.340	<0.0001
	Twitter	0.333	<0.0001
	Instagram	0.364	<0.0001
	Path	0.341	<0.0001
	MSN	0.460	<0.0001
	Sina	0.494	<0.0001
iOS Phone	Facebook	0.297	<0.0001
	Twitter	0.311	<0.0001
	Instagram	0.294	0.016
	Path	0.306	<0.0001
	MSN	0.453	<0.0001
	Sina	0.490	<0.0001
iOS Tablet	Facebook	0.299	<0.0001
	Twitter	0.299	<0.0001
	Instagram	0.297	<0.0001
	Path	0.299	<0.0001
	MSN	0.404	<0.0001
	Sina	0.389	<0.0001

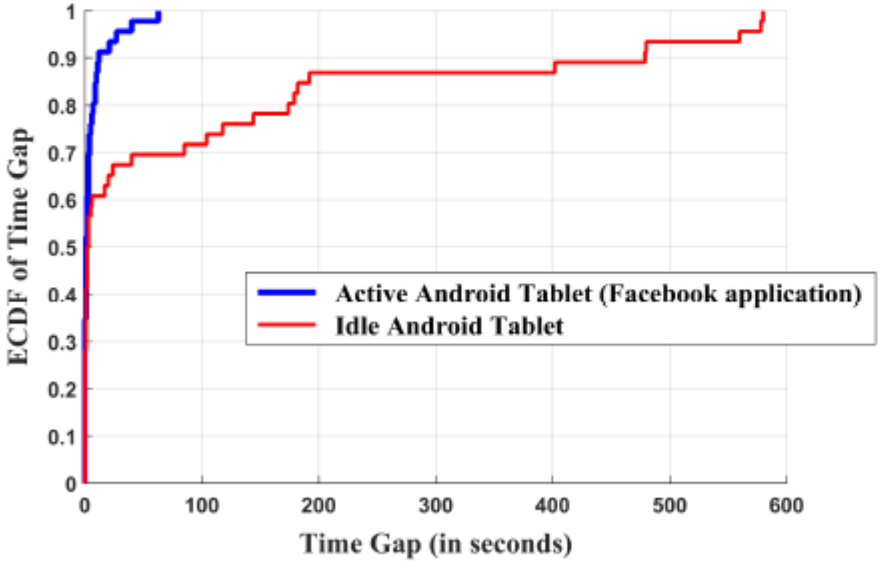


Fig. 1. ECDF of time gap feature in two cases: *Idle* VS. *Active*

Modelling

- We set *MinPts*= 3 as a fair number of URL requests in a single transaction.
  - This is based on observation that a user action on an application is often triggered by *more than one or two URL requests*.
- *Epsilon* selection:

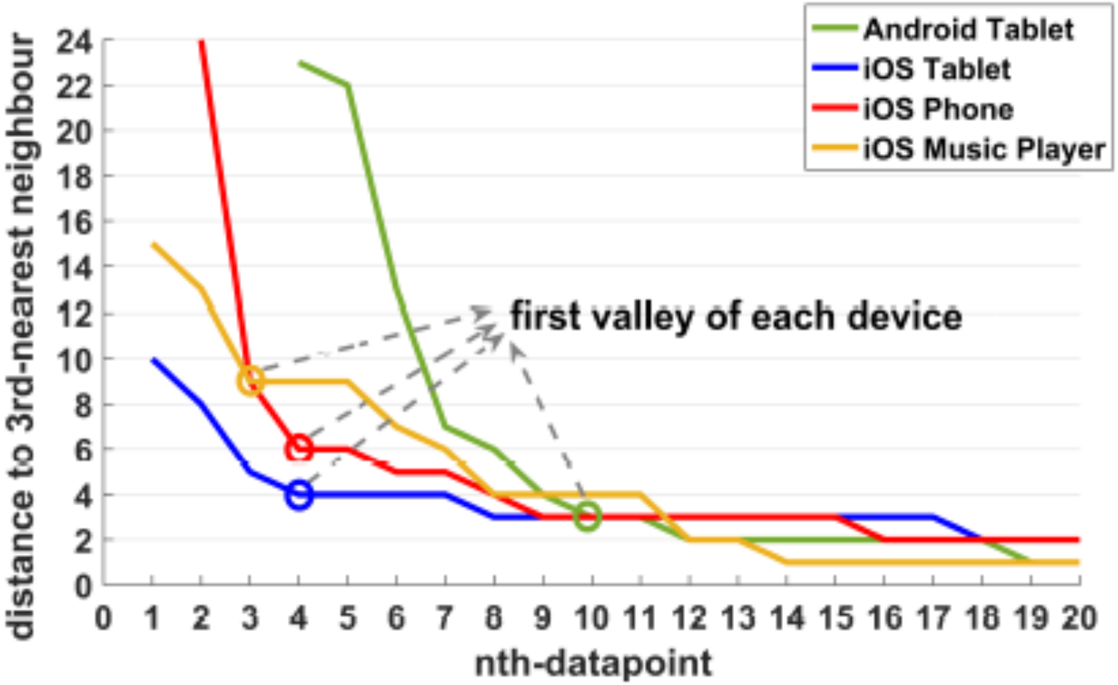
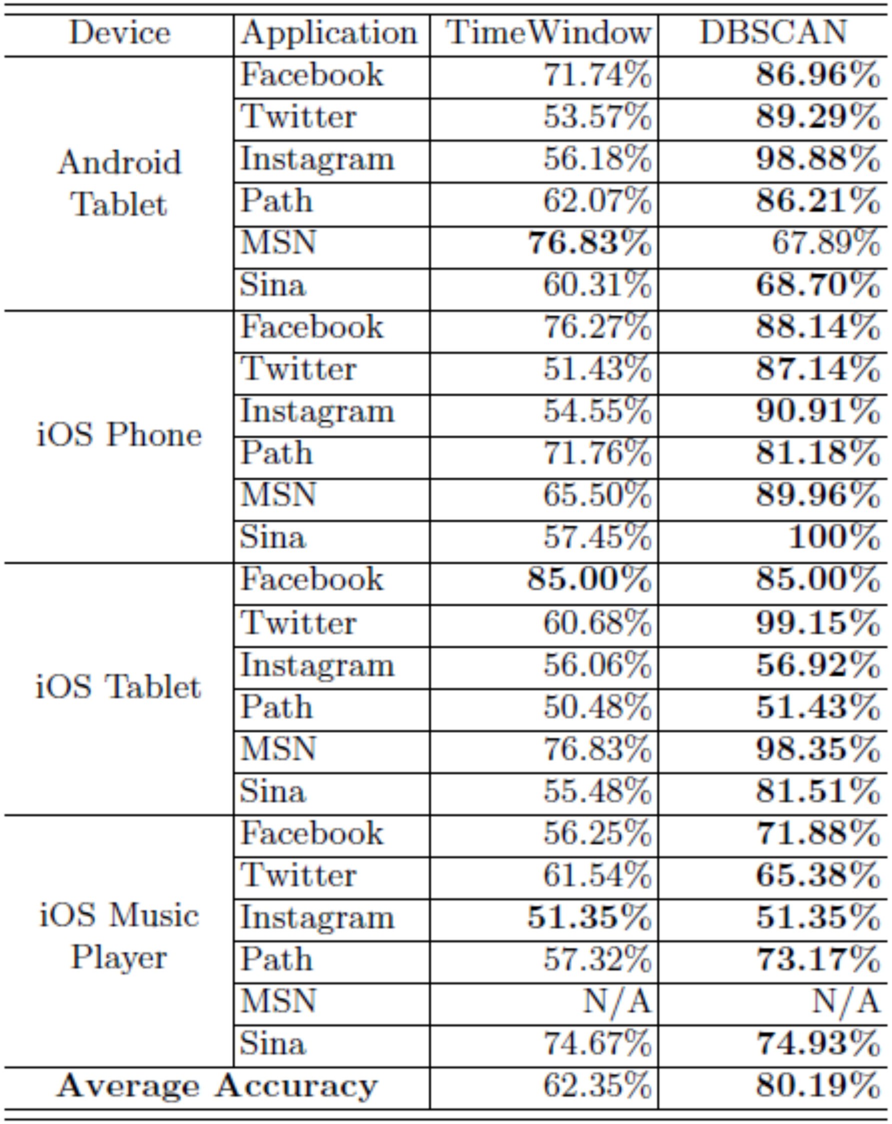


Fig. 3. Example distribution of the sorted 3rd-nearest neighbour distance

Modelling



Device	Application	TimeWindow	DBSCAN
Android Tablet	Facebook	71.74%	**86.96%**
	Twitter	53.57%	**89.29%**
	Instagram	56.18%	**98.88%**
	Path	62.07%	**86.21%**
	MSN	**76.83%**	67.89%
	Sina	60.31%	**68.70%**
iOS Phone	Facebook	76.27%	**88.14%**
	Twitter	51.43%	**87.14%**
	Instagram	54.55%	**90.91%**
	Path	71.76%	**81.18%**
	MSN	65.50%	**89.96%**
	Sina	57.45%	**100%**
iOS Tablet	Facebook	**85.00%**	**85.00%**
	Twitter	60.68%	**99.15%**
	Instagram	56.06%	**56.92%**
	Path	50.48%	**51.43%**
	MSN	76.83%	**98.35%**
	Sina	55.48%	**81.51%**
iOS Music Player	Facebook	56.25%	**71.88%**
	Twitter	61.54%	**65.38%**
	Instagram	**51.35%**	**51.35%**
	Path	57.32%	**73.17%**
	MSN	N/A	N/A
	Sina	74.67%	**74.93%**
Average Accuracy		62.35%	**80.19%**

Reference

- *Bilih Priyogi, Mark Sanderson, Flora Salim, Jeffrey Chan, Martin Tomko, Yongli Ren*.
  - **Identifying In-App User Actions from Mobile Web Logs**.
    - **PAKDD** 2018.
    - (CORE Rank A)

Data Science

Thanks!