如何奪得先機搶發 PTT 地震爆文？

2020 Spring ccClub Project

陳柏瑜、高翊傑、蔡易辰

June 24, 2020

Graduate Institute of Economics
National Taiwan University

Brief

Inroduction

Data

Results

Conclusion

Inroduction

Motivation

有使用過 PTT 的人都不陌生，每當地震後第一件要做的事情不是報平安，而是到八卦版上發地震文。

在 2020/5/3 上午 11:29，我在臺大社科院感受到地震，幾秒後我突然好奇，若在這時打開 PTT 的八卦版，會看到幾篇地震文。結果已然有十數篇，更發現第一篇早已推爆，且發文時間在 11:28:18。這意味者早在我感受到地震以前，在別處已有人感受到地震而發文。

這不但受地震的傳播速度影響，更讓我發想了以下的問題：究竟還有哪些因素影響著人們「能夠搶到 PTT 的地震爆文」呢？

問題發想

根據科學月刊的文章，P 波每秒走 5-7 km，S 波每秒走 3-4 km。而 5/3 上午的地震震央在台東臺北方海域，爆文則由來自台中的網友奪得。上一次在八卦版有文章的地震則是 4/12，震央及爆文發文者均在宜蘭。因此可以提出以下的問題：

- PTT 使用者與震央的遠近是否會顯著地影響奪得爆文的機率

與此同時，我們必須進一步問的可能就是，如果會，那麼：

- 如果其他條件不變，我所處的縣市距離震央每遠一公里，會比別人多花幾秒鐘發文？

問題發想

為了要回答這個問題，我們顯然需要發文者的位置，幸好 PTT 已經紀錄了 IP 位址，我們只要反查即可。如：反查 IP 網站

以及利用中央氣象局發佈的地震報告當中的震央資訊，即可比對發文者位置與震央位址的遠近。

此外，我們還好奇，是否有「縣市別」的因素影響著 PTT 使用者是否搶到地震爆文。例如：人口密度、各縣市的平均移動網路速度等。

Approach

我們以如下的方式獲取資料：

1. 以 Request 及 BeautifulSoup 爬 PTT 八卦版關鍵字有「地震」的文章
2. 以 Selenium 反查 IP 位址，紀錄經緯度資訊
3. 以 Pandas 套件清資料

Data

DataFrame

	groupID	Date_format	YouWin	TimeDifference	TimeDifference_Bao	Distance	ShakingExtent	InCenter	City_Translated	IsPeak	OffPeak
1	1	2019-05-08 23:49:28	1.0	41.0	0.0	24739.820786	2.0	0	NaN	1	0
2	1	2019-05-08 23:56:55	0.0	488.0	447.0	99.955575	2.0	0	臺北	1	0
5	2	2019-05-10 21:50:07	1.0	37.0	0.0	71.898385	3.0	0	臺北	0	0
8	3	2019-05-13 16:55:49	1.0	28.0	0.0	49.557932	3.0	0	臺北	1	0
9	3	2019-05-13 16:55:51	0.0	30.0	2.0	49.557932	3.0	0	臺北	1	0
10	3	2019-05-13 16:55:54	0.0	33.0	5.0	49.557932	3.0	0	臺北	1	0
11	3	2019-05-13 16:55:59	0.0	38.0	10.0	49.557932	3.0	0	臺北	1	0
12	3	2019-05-13 16:56:02	0.0	41.0	13.0	49.557932	3.0	0	臺北	1	0
13	3	2019-05-13 16:56:04	0.0	43.0	15.0	49.557932	3.0	0	臺北	1	0
14	3	2019-05-13 16:56:10	0.0	49.0	21.0	49.557932	3.0	0	臺北	1	0
15	3	2019-05-13 16:56:21	0.0	60.0	32.0	49.557932	3.0	0	臺北	1	0

Table of Variables

Variable	Description	Type	單位
YouWin	是否是地震爆文	binary	
TimeDifference	文章發表時間與地震發生時間的時間差	continuous	Second
TimeDifference_Bao	地震文與地震爆文	continuous	second
Distance	發文者與震央的距離	continuous	km
ShakingExtent	震央所在地的震度	level	
InCenter	發文者是否為在震央所在的縣市	binary	
IsPeak	發文者是否在尖峰時段（22:00～2:00;13:00～17:00）發文	binary	
OffPeak	發文者是否在離峰時段（4:00～8:00; 18:00～20:00）發文	binary	
PopDensity	各縣市人口密度	continuous	人/平方公里
download_4G	各縣市 4G 移動網路平均下行速率	continuous	Mbps
upload_4G	各縣市 4G 移動網路平均上行速率	continuous	Mbps

About Cleaning Data

由於由 IP 位址反查的經緯度僅能定位到「縣市」層級的地理位置，因此與縣市別有關的資料，如：人口密度、4G 移動網路下行、上行速率等資料都是以縣市別為單位。

注意到，我們考量的是一場地震的震度，而非規模。因為我們相信震度更能體現體感，即震度能考量到同樣規模下，深、淺層地震給人感受的差別。又，事實上震度應因縣市而有所不同，然而該資料不易抓取，因此僅納入震央所在地的震度。

About Cleaning Data

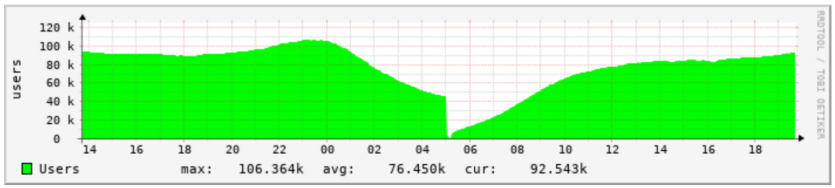
尖峰時段與離峰時段的定義方式參考了PTT Statistics的統計數據

將波峰的前後兩小時定義為尖峰時段；將波谷的前後兩小時定義為離峰時段。

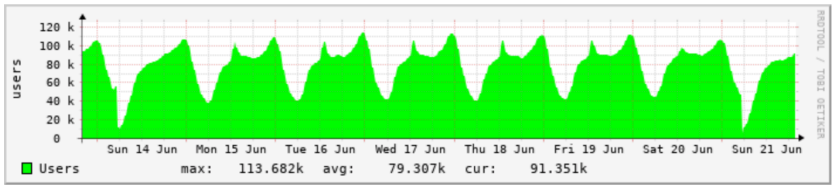
Online Users for PTT

The statistics were last updated **Sun Jun 21 19:40:14 2020**

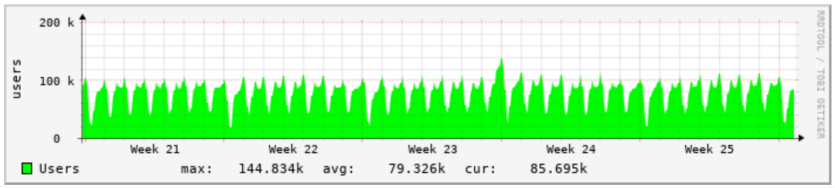
`Daily' Graph (5 Minute Average)



`Weekly' Graph (30 Minute Average)



`Monthly' Graph (2 Hour Average)



About Cleaning Data

各縣市的 4G 移動網路速度來自NCC 發佈的報告。

表 7、移動量測之各縣市平均行動上網速率（單位：Mbps）

平均速率 / 縣市	4G 平均下載速率	4G 平均上傳速率
基隆市	58.66	17.21
宜蘭縣	39.63	12.30
臺北市	56.01	16.41
新北市	47.16	13.74
桃園市	48.78	13.94
新竹市	51.53	15.00
新竹縣	51.08	13.74
苗栗縣	69.03	16.37
臺中市	65.02	20.52
彰化縣	62.60	17.19
南投縣	59.20	13.81
雲林縣	59.73	17.22
嘉義市	62.62	20.01
嘉義縣	53.32	13.09
臺南市	62.18	15.17
高雄市	53.25	15.73
屏東縣	54.42	14.10
臺東縣	60.66	13.18
花蓮縣	43.10	13.12
金門縣	68.20	11.35
澎湖縣	65.54	16.37

About Cleaning Data

各縣市人口密度資料來自Wikipedia，由臺灣行政區面積表及臺灣行政區人口列表的資料計算出。

與震央間的距離由經緯度之差簡單換算而得：

AUG 01 WED 2007 09:00

台灣地區經緯度的距離

24°N, 121°E 位於南投埔里西關刀山附近

如果以 24°N, 121°E 為原點

原點往	經緯度	至	兩點距離(公尺)
西	1°	24°N,120°E	101751.561277
東	1°	24°N,122°E	101751.561277
南	1°	23°N,121°E	110751.075273
北	1°	25°N,121°E	110765.515243

Results

Models

我們分別以 `YouWin`, `TimeDifference`, `TimeDifference_Bao` 為被解釋變數。

我們分別以以下三個模型：

1. Binary Response
2. Linear Model
3. Poisson Regression

來看距離 (Distance) 對上述被解釋變數的效果為何。

Regression Table: Binary Response

VARIABLES	(1) LPM:POLS	(2) LPM:RE	(3) LPM:FE	(4) Probit	(5) Logit
distance	-0.00124***	-0.000957***	-0.000296	-0.00416***	-0.00728***
	(0.000407)	(0.000348)	(0.000500)	(0.00150)	(0.00275)
shakingextent	0.0195	-0.00609		0.0645	0.104
	(0.0397)	(0.0391)		(0.124)	(0.218)
incenter	0.00297	0.102	0.201	-0.0450	-0.0844
	(0.177)	(0.168)	(0.205)	(0.502)	(0.777)
ispeak	-0.0732	-0.0434		-0.209	-0.360
	(0.107)	(0.0753)		(0.314)	(0.516)
offpeak	-0.183*	-0.218**		-0.640**	-1.100**
	(0.0955)	(0.0890)		(0.297)	(0.484)
popdensity	-2.81e-05***	-1.85e-05**	-1.40e-05*	-8.48e-05***	-0.000145***
	(8.49e-06)	(7.45e-06)	(7.23e-06)	(2.51e-05)	(4.20e-05)
download_4g	-0.0141	-0.0136	-0.0184	-0.0420	-0.0693
	(0.0113)	(0.0110)	(0.0162)	(0.0330)	(0.0540)
upload_4g	0.0343	0.0396	0.0483	0.111	0.189
	(0.0286)	(0.0277)	(0.0373)	(0.0821)	(0.137)
Constant	0.847*	0.872*	0.619	0.957	1.555
	(0.508)	(0.472)	(0.528)	(1.470)	(2.309)
Observations	336	336	336	336	336
R-squared	0.143		0.096		
Number of groupid		87	87		

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

APE for Probit

```
Marginal effects after probit
      y  = Pr(youwin) (predict)
         =  .25759779
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.      z    P>|z|     [    95% C.I.   ]       X
---------+--------------------------------------------------------------------
distance |  -.0013423      .00047    -2.83   0.005    -.002273   -.000412   103.467
shakin~t |   .0208306      .03972     0.52   0.600    -.05702    .098681    3.11012
incenter*|  -.0143652       .1585    -0.09   0.928    -.325023   .296293    .136905
  ispeak*|  -.0659249      .09777    -0.67   0.500    -.257548   .125698    .357143
 offpeak*|  -.1906436      .08894    -2.14   0.032    -.364967   -.01632     .33631
popden~y |  -.0000274      .00001    -3.61   0.000    -.000042   -.000013   6288.73
downl~4g |  -.0135642      .01079    -1.26   0.209    -.034709    .00758    56.3185
uploa~4g |   .0358601      .02718     1.32   0.187    -.01742    .08914    16.3576
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

APE for Logit

```
Marginal effects after logit
      y  = Pr(youwin) (predict)
         =  .24935992
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.      z    P>|z|     [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
distance |   -.001363     .00049    -2.80   0.005    -.002318   -.000408   103.467
shakin~t |    .0195255    .04058     0.48   0.630    -.060004    .099055    3.11012
incenter*|   -.0155464    .14083    -0.11   0.912    -.291576    .260483    .136905
  ispeak*|   -.0655836    .09196    -0.71   0.476    -.245821    .114653    .357143
 offpeak*|   -.1867486    .08456    -2.21   0.027    -.352486   -.021012     .33631
popden~y |   -.0000271    .00001    -3.84   0.000    -.000041   -.000013   6288.73
downl~4g |   -.0129651    .01029    -1.26   0.208    -.033138    .007208   56.3185
uploa~4g |    .0354114    .02653     1.33   0.182    -.016588    .087411   16.3576
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Interpretations: Binary Response

在控制住個別縣市的因素後，我們能發現距離 (Distance) 在絕大部分的模型都能影響「搶到爆文」的機率。

然而其效果大約為：在其他條件不變下，每離震央遠一公里，搶到爆文的機率就下降 0.13%

值得一提的是：一縣市人口密度越高，可能表示網路使用者之間的競爭越激烈，因此與搶到爆文的機率有負向關係。然而人口密度 (Density) 是控制變數，並非 variable of interest，我們暫且不討論其可能存在內生性的問題。

Regression Table: TimeDifference

VARIABLES	(1) POLS	(2) RE	(3) FE	(4) Poisson
distance	0.495***	0.495***	0.267	0.00361***
	(0.133)	(0.133)	(0.230)	(0.000844)
shakingextent	8.619	8.619		0.0614
	(9.050)	(9.050)		(0.0685)
incenter	31.71	31.71	-8.303	0.215
	(24.31)	(24.31)	(52.53)	(0.218)
ispeak	-1.807	-1.807		0.00103
	(24.67)	(24.67)		(0.196)
offpeak	-22.71	-22.71		-0.155
	(21.82)	(21.82)		(0.171)
popdensity	0.00612***	0.00612***	0.00326	5.38e-05***
	(0.00185)	(0.00185)	(0.00276)	(1.58e-05)
download_4g	2.380	2.380	1.788	0.0230
	(1.766)	(1.766)	(3.341)	(0.0158)
upload_4g	-4.095	-4.095	-0.423	-0.0369
	(4.795)	(4.795)	(7.375)	(0.0454)
Constant	-55.06	-55.06	-16.19	3.211***
	(68.02)	(68.02)	(123.6)	(0.631)
Observations	336	336	336	336
R-squared	0.084		0.030	
Number of groupid		87	87	

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

APE for Poisson

```
Marginal effects after poisson
      y  = Predicted number of events (predict)
         =  119.31082
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.      z    P>|z|     [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
distance |    .4306986      .10181     4.23   0.000     .231162    .630235   103.467
shakin~t |    7.326489     8.07731     0.91   0.364    -8.50474    23.1577   3.11012
incenter*|     27.7357      30.468     0.91   0.363    -31.9805    87.4519   .136905
  ispeak*|    .1230213      23.418     0.01   0.996    -45.7761    46.0221   .357143
 offpeak*|   -18.04152      19.644    -0.92   0.358    -56.5427    20.4597    .33631
popden~y |    .0064178      .00192     3.35   0.001     .002661    .010175   6288.73
downl~4g |    2.741332     1.87908     1.46   0.145    -.941589    6.42425   56.3185
uploa~4g |   -4.400671     5.36889    -0.82   0.412    -14.9235    6.12216   16.3576
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Interpretations: Time Difference

在控制住個別縣市的因素後，我們能發現距離 (Distance) 在「地震後發文的時間差」的 APE 約為 0.43 秒，即「每離震央遠一公里，就需要多 0.43 秒來發文」。

其他值得一看的有：

1. 在人口密度越高的城市，需要越多時間來發地震文。
2. 4G 上行速率每增加 1Mbps，就可以減少 4.4 秒鐘的發文時間。

同樣地，我們暫且忽略 `Distance` 以外變數的一致性問題。

Regression Table: TimeDifference between 爆文 &Others

VARIABLES	(1) POLS	(2) RE	(3) FE	(4) Poisson
distance	0.449***	0.449***	0.312	0.00483***
	(0.122)	(0.122)	(0.218)	(0.00114)
shakingextent	14.29	14.29		0.149
	(10.13)	(10.13)		(0.110)
incenter	34.28	34.28	0.747	0.344
	(27.49)	(27.49)	(50.45)	(0.358)
ispeak	7.153	7.153		0.115
	(25.11)	(25.11)		(0.291)
offpeak	-41.77*	-41.77**		-0.493*
	(21.28)	(21.28)		(0.269)
popdensity	0.00576***	0.00576***	0.00327	8.18e-05***
	(0.00182)	(0.00182)	(0.00264)	(2.64e-05)
download_4g	1.862	1.862	1.837	0.0313
	(1.741)	(1.741)	(3.172)	(0.0256)
upload_4g	-2.828	-2.828	-0.782	-0.0553
	(4.503)	(4.503)	(7.018)	(0.0754)
Constant	-99.96	-99.96	-64.62	2.006*
	(76.11)	(76.11)	(121.0)	(1.073)
Observations	336	336	336	336
R-squared	0.086		0.032	
Number of groupid		87	87	

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

APE for Probit & Logit

```
Marginal effects after poisson
      y  = Predicted number of events (predict)
         =  71.157265
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.      z    P>|z|   [     95% C.I.   ]      X
---------+--------------------------------------------------------------------
distance |   .3433456      .07917     4.34   0.000    .188182   .498509   103.467
shakin~t |   10.63704     7.70775     1.38   0.168   -4.46987    25.744   3.11012
incenter*|   27.88377       32.74     0.85   0.394   -36.2848   92.0524   .136905
  ispeak*|   8.350337      21.136     0.40   0.693   -33.0764   49.7771   .357143
 offpeak*|  -32.70711      17.594    -1.86   0.063   -67.1907   1.77644    .33631
popden~y |   .0058235      .00182     3.20   0.001    .002255   .009392   6288.73
downl~4g |   2.225976     1.82652     1.22   0.223   -1.35394   5.80589   56.3185
uploa~4g |   -3.93216      5.3404    -0.74   0.462   -14.3991   6.53483   16.3576
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Interpretations: Time Difference between 爆文 & Others

由於中央氣象局發佈的地震報告，其內紀錄的地震發生時間系回推的地震發生時間。因此即便是「地震爆文」，其與地震發生時間的時間差亦在 30 至 40 秒。又因此，我們好奇「搶發爆文失敗者」跟「成功者」之間的時間差作為被解釋變數時，各個因素的效果大小。

Interpretations: Time Difference between 爆文 & Others

在控制住個別縣市的因素後，我們能發現距離 (Distance) 在「地震後發文的時間差」的 APE 變小至 0.34 秒，即「每離震央遠一公里，就需要多 0.34 秒來發文」。

其他值得一看的有：

1. 在離峰時段發文，需要用的時間較少
2. 在人口密度越高的城市，需要越多時間來發地震文。
3. 4G 上行速率每增加 1Mbps，就可以減少 3.9 秒鐘的發文時間。

同樣地，我們暫且忽略 `Distance` 以外變數的一致性問題。

Conclusion

Questions Asked

1. PTT 使用者與震央的遠近是否會顯著地影響奪得爆文的機率
   - 會
2. 如果其他條件不變，我所處的縣市距離震央每遠一公里，我搶到爆文的機率會低多少？
   - 大約 0.13%
3. 如果其他條件不變，我所處的縣市距離震央每遠一公里，會比別人多花幾秒鐘發文？
   - 大約 4 秒鐘

Possible Approaches

這份專題研究仍有不足之處，或可朝以下方向改進：

1. 累積更長期間的資料，以避免少數縣市幾乎沒有 observation 的問題
2. 加入每起地震在各縣市的震度資料作為控制變數
3. 少部分由 IP 定位至「鄉鎮」層級的地理資訊受限於絕大部分資料只能定位至「縣市」層級，無法得出其細節資訊

視覺化

針對此次專案，我們所整理的資料結合地圖來視覺化。