P-E-R-S-I-S-T-E-N-C-E and DISTINCTIVENESS of Inter-event Time Distributions in Online Human Behavior

Jiwan Jeong and Sue Moon
School of Computing, KAIST

In TempWeb ‘17 (WWW ’17 Companion)
April 3, 2017
What is inter-event time?

• Time gap between two consecutive events
• E.g., earthquake waves, packet arrivals, ...
Our definition of inter-event time

• Time gap between two consecutive *actions* in a service by one person
• E.g., tweeting, blog posting, email sending, ...

• Simply put
  • Inter-event time = interval
  • Inter-event time distribution = interval pattern
Previous studies focused on

• Characterizing **aggregate** interval patterns
  • Web re-visit pattern [Adar *CHI* 2007][Adar *CHI* 2008]
  • Web browsing pattern [Kumar *WWW* 2010]
  • Service usage pattern [Halfaker *WWW* 2015]

• Finding **universal laws** among interval patterns
  • Power-law by priority queuing process [Barabasi *Nature* 2005]
  • Log-normal by non-homogeneous Poisson process [Malmgren *PNAS* 2008]
We focus on individual-level

• How does an individual’s interval pattern change over time?
• Does it remain consistent or fluctuate from time to time?
• How distinctive is it from those of others?
Individuals have interval patterns that are persistent over time, but distinctive from others.
Tweets by Ellen DeGeneres

Twitter timeline
@TheEllenShow

2014-12-05
2015-06-16
2015-11-02
2016-05-11
2016-10-02

Tweeting Time

Density

Inter-tweeting Time

[Graph showing tweeting and inter-tweeting times with labels and axes]
Tweets by Jimmy Fallon
Tweets by Sue Moon

@sbmoon

Density

Tweeting Time

Inter-tweeting Time
Tweets by **Albert-László Barabási**

![Graph showing the distribution of tweeting times and inter-tweeting times for different periods.](image-url)
Tweets by Eytan Adar
Tweets by Aaron Clauset
Tweets by Nicolas Christakis
Tweets by Alex Vespagini
Tweets by Andrew Ng
Tweets by Ed Chi

@edchi

Tweeting Time

Inter-tweeting Time
Tweets by Bruno Gonçalves
Tweets by Haewoon Kwak

@haewoon

2013-02-01

2015-03-19

2015-10-26

2016-05-25

2016-10-01

TWEETING TIME

INTER-TWEETING TIME
Tweets by Carlos Castillo

@ChaToX

![Graph showing tweeting and inter-tweeting times](image-url)
Tweets by Peter Dodds

@peterdodds

Tweeting Time

Inter-tweeting Time
In this work

• Design a computation framework to quantify interval patterns
• Show their persistence and distinctiveness
• Use interval patterns to distinguish one user from others
Datasets for this study

- Wikipedia: 15 years of entire history
- me2day: 7 years of entire history
- Twitter: 3000 recent tweets per user
- ENRON: 3 years of email history
Estimate interval patterns

Compare interval patterns

Design computation framework
Estimate interval patterns

Compare interval patterns

Design computation framework
Convert **discrete intervals** to **continuous PDF**
Gaussian kernel density estimation

For multi-modal distributions, we use Sheather and Jones’ bandwidth
Now, we can estimate interval patterns!
Estimate interval patterns

Compare interval patterns

Design computation framework
Calculate **distance** between interval patterns
Jensen-Shannon distance

- A **metric** of the difference between probability density functions
  - Non-negative: $d(x, y) \geq 0$
  - Identity of indiscernibles: $d(x, y) = 0$ iff $x = y$
  - Symmetry: $d(x, y) = d(y, x)$
  - Subadditivity: $d(x, z) \leq d(x, y) + d(y, z)$
Now, we can compare interval patterns!
Estimate interval patterns

Compare interval patterns

Design computation framework
Define **self-distance** and **reference distance**
Experimental settings for longitudinal analysis

- Select users with +500 actions on each service
- Divide each user’s timeline into 10 windows

\[
\begin{array}{cccccc}
W_1 & W_2 & \ldots & W_9 & W_{10} \\
\end{array}
\]

- \( \binom{10}{2} \) = 45 self-distances for each user
- 10 × 10 = 100 reference distances for each pair of users
PERSISTENCE
&
DISTINCTIVENESS
Persistence and distinctiveness are relative

- If $d^{self}$ are small, the pattern is persistent
- How small should it be?
- If $d^{self} < d^{ref}$, the pattern is persistent [Saramäki PNAS 2014]

- Furthermore, if $d^{self} \ll d^{ref}$, the patterns are distinctive
$d^\text{self} \, \text{vs} \, d^\text{ref}$
How long do interval patterns persist?

• Binning $d^{\text{self}}$ by the **time gap** between two windows

• Compare binned $d^{\text{self}}$ with overall $d^{\text{ref}}$
Persistence over time

Binned into 6 groups
Persistence over time

me2day

0.1 0.2 0.3 0.4 0.5 0.6

d^{self} d^{ref} 1w 1mo 3mo 1yr 3yr
Persistence over time

Enron Email

![Box plot showing persistence over time with different time intervals: d\text{self}, d\text{ref}, 1w, 1mo, 3mo, 1yr, 3yr.](image)
Do interval patterns persist after long inactivity?

- Binning $d^{\text{self}}$ by the **longest interval** between two windows

- Compare binned $d^{\text{self}}$ with overall $d^{\text{ref}}$
Persistence after inactivity

[Box plot diagram showing persistence levels over time]
Persistence after inactivity
Do interval patterns persist through changing daily routine?

- Binning $d_{\text{self}}$ by the **circadian distance** between two windows

```
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Wi</td>
<td></td>
<td>Wj</td>
</tr>
</tbody>
</table>
```
Persistence through changing daily routine
In summary,

• Individuals have *interval signatures* that *persist over years*
• The signatures persist *even after coming back from long inactivity*
• The signatures persist *through changing daily routine*
User Identification Using *Interval Signatures*
User identification: Problem definition

• Given two windows each containing 100 intervals

![Diagram of W_A and W_B]

• Can we determine those from the same user or not?
A very simple identifier

Calculate the distance $d$

If $d < \text{threshold}$,

Else,
### Identification performance \((1 - \text{Equal Error Rate})\)

<table>
<thead>
<tr>
<th></th>
<th>Wikipedia</th>
<th>me2day</th>
<th>Twitter</th>
<th>Enron</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consecutive</td>
<td>80%</td>
<td>87%</td>
<td>83%</td>
<td>76%</td>
</tr>
<tr>
<td>&gt; 1 year gap</td>
<td>71%</td>
<td>78%</td>
<td>76%</td>
<td>71%</td>
</tr>
</tbody>
</table>

- Performance of other behavioral biometrics
  - Keystroke dynamics: $\sim90\%$ [Peacock IEEE S&P 2004]
  - Mouse dynamics: $\sim80\%$ [Jorgensen AsiaCCS 2011]
  - Gaits: $\sim80\%$ [Gaufrov University of Oslo 2008]
Follow-up questions

• What do people with similar interval signatures have in common?
• What can be inferred about users by analyzing interval signatures?
• How interval signatures are related to other personal characteristics?
Interval Signature: P-E-R-S-I-S-T-E-N-C-E and DISTINCTIVENESS of Inter-event Time Distributions in Online Human Behavior

Q&A
## Dataset statistics

<table>
<thead>
<tr>
<th># of users</th>
<th>Wikipedia</th>
<th>me2day</th>
<th>Twitter</th>
<th>Enron</th>
</tr>
</thead>
<tbody>
<tr>
<td>With &gt;25 actions</td>
<td>521K</td>
<td>587K</td>
<td>921K</td>
<td>937K</td>
</tr>
<tr>
<td>With &gt;100 actions</td>
<td>165K</td>
<td>203K</td>
<td>768K</td>
<td>542K</td>
</tr>
<tr>
<td>With &gt;500 actions</td>
<td>47K</td>
<td>43K</td>
<td>334K</td>
<td>65K</td>
</tr>
</tbody>
</table>
$d^{self}$ vs $d^{ref}$ at different window sizes
K-means clustering of interval patterns
Joint probability matrix for transition $\mathcal{W}_i \rightarrow \mathcal{W}_{i+1}$