# Interval Signature: Persistence and Distinctiveness of Inter-event Time Distributions in Online Human Behavior

Jiwan Jeong
School of Computing
KAIST
jiwanjeong@gmail.com

Sue Moon
School of Computing
KAIST
sbmoon@kaist.edu

## ABSTRACT

*Interval patterns*, or inter-event time distributions that occur in human activity, have long been an interest of many researchers studying human dynamics. While previous studies have mostly focused on characterizing the aggregated inter-arrival patterns or finding universal patterns across all individuals, we focus on the diversity among the patterns of different individuals; the goal of this paper is to understand how persistent an individual's interval pattern is and how distinctive it is from those of the others. We use Wikipedia, me2DAY, Twitter, and Enron email data to study the interval patterns of online human behavior. Our analysis reveals that individuals have robust and unique *interval signatures*. The interval pattern of a user tends to persist over years, even after coming back from a long hiatus of inactivity, despite considerable change in circadian rhythms. Furthermore, the interval patterns of individuals are highly distinct from that of others. We put our new findings in practical use of identifying users.

## Keywords

Inter-event time; human behavior; personal characteristic; behavioral signature; interval signature;

## 1. INTRODUCTION

Vibrant human dynamics manifests in rich patterns and leads to complex *inter-event times* of individual human actions. We access numerous online services of this 21st century countless times a day and leave traces of our activities online. Let us take Facebook as an example. An individual may update Facebook status very frequently, say, every other minute, or only daily and weekly, or only after a long hiatus of a few months. If we capture the inter-event time of a user's activities in a distribution, what does the distribution tell us?

Previous studies have mostly focused on analyzing aggregated inter-arrival patterns from the perspective of a service provider [1, 2, 10, 20, 42], or finding universal patterns across all individuals in order to explicate or model the human behavior [3, 13, 17, 25, 26, 36, 41]. On the contrary, we take another angle and examine the diversity in individual distributions of inter-event times, which we call *interval pattern* in this work. How does an individual's interval pattern change over time? Does it remain consistent or fluctuate from time to time? If the pattern is persistent, is it distinctive enough to characterize one from the others?

In this work we use four online datasets—Wikipedia edit history, me2DAY, Twitter, and Enron email—to study the interval patterns of users across different platforms. In particular, we have the entire history of Wikipedia and me2DAY, which enables us to study longitudinal changes over the years in individuals' interval patterns.

Our analysis reveals that individuals have persistent and distinct interval patterns in all four platforms. For a given user, the interval pattern persists over years, even after coming back from a long hiatus of inactivity or despite considerable change in circadian rhythms. Even more, the interval pattern of an individual is highly distinct from those of others. Also, we show that abrupt changes in interval patterns are mostly due to instant bursts rather than changes in persisting trends. This result implies that the interval pattern is a coherent personal characteristic.

Based on our findings, we use interval patterns in user identification and linking an individual's accounts. Our prototype implementation demonstrates that the interval pattern is a good behavior feature distinguishing one from the others. In this era of digital footprints, our findings from this work have far-reaching consequences. We show that an interval pattern is representative of a person, distinguishing one from the rest, and adds a new dimension in capturing one's personality or preference.

## 2. DEFINING TERMS WITH SAMPLES

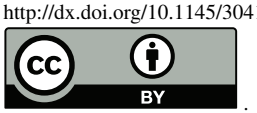
Figure 1 displays Twitter timelines and the longitudinal inter-tweeting time distributions for three famous computational social scientists, Lada Adamic, Albert-Laszlo Barabasi, and Nicolas Christakis. In addition, on our project web page[1] we uploaded more samples from many other researchers and 10 celebrities who actually run their own Twitter accounts [11]. In this section, we introduce the terms we use throughout this paper, using these samples.

### 2.1 Intervals, Windows, and the Patterns

An *interval* or *inter-event time*, $\tau$, is the time gap between two consecutive actions by the same individual on a single

[1] https://j1wan.github.io/interval-signature/


*WWW 2017 Companion,* April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3041021.3051115



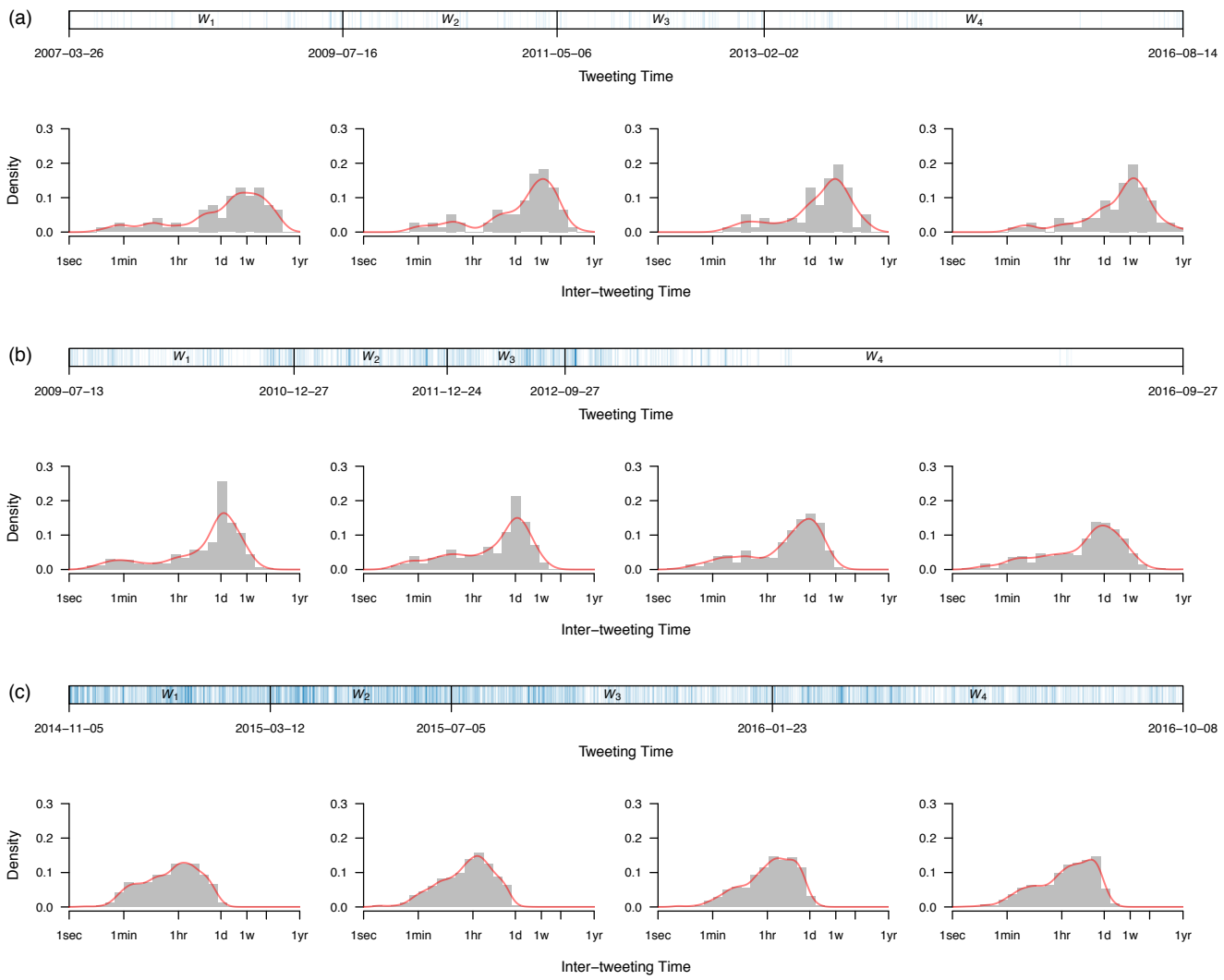



**Figure 1: Longitudinal changes in tweeting interval patterns for famous computational social scientists: (a) Lada Adamic, (b) Albert-Laszlo Barabasi, and (c) Nicolas Christakis.**

online service. In timelines in Figure 1, each vertical line represents when the tweet was created, and a length between two neighboring vertical lines represents an interval.

A *window* is a time period containing a set of consecutive intervals. We refer to the number of intervals in a window as the *size* and the time duration as the *length*. For example, in Figure 1, we set each window, from $W_1$ to $W_4$, to have $\lfloor (n-1)/4 \rfloor$ intervals where $n$ is the total number of tweets written by each user. Their lengths in time differ from one another. In the sample figures, we split the windows to have an equal number of intervals rather than an equal length in time. That is, we fix the size of the windows constant. If we set the windows to be of fixed length in time, some windows would contain zero or only a few intervals during the period of low activity. We want all the windows to have a sufficient number of samples.

The *interval pattern* of a window is its inter-event time distribution in log-scale. We transform the discrete time intervals to a continuous probability density function using Gaussian kernel density estimation with Sheather and Jones' bandwidth selector [15, 31]. (See Appendix A.1 for detail.) With providing evidence that each individual's interval pattern is persistent and distinctive, we will call the shape as *interval signature*.

## 2.2 Distances between Patterns and Users

A noteworthy observation in the samples is the consistency in each individual's interval patterns. Throughout years of time, the shape of interval patterns of a single user remains consistent over windows, while the shapes are quite different from each other. Then, how can we quantify the self-consistency and cross-individual diversity?

We begin by measuring the distance between two interval patterns. Given any two windows, we calculate the distance between those interval patterns using Jensen-Shannon distance [7, 19, 21]. (See Appendix A.2 for detail.)

Then, we call the distance between a user's two different windows as *self-distance*, $d^{\text{self}}$. The smaller the self-distance

is, the more consistent the individual's interval pattern is. In Figure 1, each user has 4 windows, thus we have $\binom{4}{2} = 6$ self-distances for each user; the average self-distance, $\langle d^{\text{self}} \rangle$, is $0.13 \pm 0.02$, $0.11 \pm 0.03$, and $0.12 \pm 0.04$ for Adamic, Barabasi, and Christakis, respectively. For the comparison, we refer the distance between different users' windows as *reference distance*, $d^{\text{ref}}$. In Figure 1, each pair of users has $4 \times 4 = 16$ reference distances, and the average of them, $\langle d^{\text{ref}} \rangle$, is $0.36 \pm 0.04$, $0.39 \pm 0.05$, and $0.57 \pm 0.03$ for each pair of Adamic-Barabasi, Barabasi-Christakis, and Christakis-Adamic, respectively. As the self-distances are conspicuously less than the reference distances, we distinguish a user's interval pattern from the others[2].

So far, we look at the individual-level interval patterns for dozens of sample users. Now we conduct analysis on more datasets to study longitudinal interval patterns at a population-level.

# 3. DATASET DESCRIPTION

To study *interval patterns* in various types of online behavior, we use the following datasets: Wikipedia revision history, me2DAY, Twitter, and Enron email. In particular, Wikipedia and me2DAY datasets contain the complete updates that occurred since the service launch, enabling longitudinal studies of each and every user. Table 1 summarizes the datasets, and the following subsections briefly describe each dataset.

| Number of users | Wiki | me2DAY | Twitter | Enron |
|---|---|---|---|---|
| with >25 actions | 521K | 587K | 921K | 937 |
| with >50 actions | 297K | 356K | 768K | 542 |
| with >100 actions | 165K | 203K | 624K | 298 |
| with >500 actions | 47K | 43K | 334K | 65 |

**Table 1: Dataset Statistics. Here, the *action* means platform-specified activity: article editing, posting, tweeting, and email sending for each platform.**

## 3.1 Wikipedia

Wikipedia is an ideal dataset for this study as it has long history spanning 15 years, and provides the entire edit history dump. We use the English Wikipedia revision logs[3] up to 2015. As our study focuses on human behavior, we exclude all bots [37] and blocked accounts [38] that are officially listed by Wikipedia. We assume all the other accounts as personal accounts, as Wikipedia's policy [39, 40] prohibits sharing an account between more than one individual.

## 3.2 me2DAY

A Korean microblogging and social networking service, me2DAY, was launched in February 2007 and closed in June 2014. It had a 150-character of length limitation per post as in Twitter and allows to write comments and click 'metoo' to a post as in Facebook. In May 2014, on the news of its impending closure, we crawled the entire history of me2DAY, thanks to me2DAY imposing no limit on third-party crawling. We could crawl the entire posts of each and every individual.

For our previous work [12], we cleaned the me2DAY data and labeled spam-suspicious accounts analyzing the contents. We excluded these accounts from our main analysis.

## 3.3 Twitter

Twitter, with its openness, has become probably the most popular dataset of this decade in social media studies. Using the Twitter API, we took a subset of users from Twitter and collected the tweets they have written. We used snowball sampling with two Korean celebrities as the seed set. Due to Twitter API rate limits, we collected up to 3,200 most recent tweets per user up until March, 2014.

## 3.4 Enron

For interval patterns in email behavior, we use Enron dataset [18]. The dataset contains emails from September 1998 to September 2002. We remove duplicates, and group the emails by senders. Also, we exclude non-individual senders such as announce@enron.com or support@enron.com.

# 4. LONGITUDINAL INTERVAL PATTERNS

In section 2, we split each user's timeline into 4 windows for visualization. In this section, however, we split each user's timeline into smaller pieces to observe the longitudinal interval pattern changes.

In order to make each user have enough windows and activities for longitudinal study, we only consider the users with more than 500 actions and split each user's timeline into 10 windows; each user has 10 windows having 50 or more intervals[4]. The number of those users in each platform is shown in Table 1.

We first split each user's activity history into 10 consecutive *windows*, $W_1, W_2, ..., W_{10}$, so that each window contains exactly $\lfloor (n-1)/10 \rfloor$ intervals where $n$ is the total number of actions. This window setting has the following advantages. First, in an online service, individuals have different lifespans; some users abandon a service only a few days after joining, while others remain active for several years. Segmenting the lifespans of users into equal number of windows allows us to compare individuals' longitudinal interval patterns in a single framework, even when the users have different lifespans [6]. Second, having 10 windows per each users, we can compare 45 window pairs evenly covering each users lifespan with various time gap. Using the enough number of window pairs, we can compare the self-distances under various conditions.

In this setting, each user has 45 self-distances and each pair of users generates 100 reference distances. We plot the distribution of all the distances in Figures 2(a-d) vertically aligned. These plots serve as a ruler for the distances presented in the following subsections.

## 4.1 Over user lifespan

How individuals' interval patterns change from their joining to abandoning the community? To see how the users' interval patterns change over their lifespans, we display the self distance between two consecutive windows $W_i$ and $W_{i+1}$, in Figures 2(e-h). In all four platforms, Wikipedia, me2DAY,

[2]The method that comparing self-distance and reference distance is developed upon the social signature work [30].

[3]http://dumps.wikimedia.org/enwiki/

[4]We also conducted same analysis with all the users with more than 100 actions splitting each timeline into 4 windows. In addition, we conducted analysis with fixed window sizes. In all cases, the results were consistent. (For the window size effect, see Appendix B.)

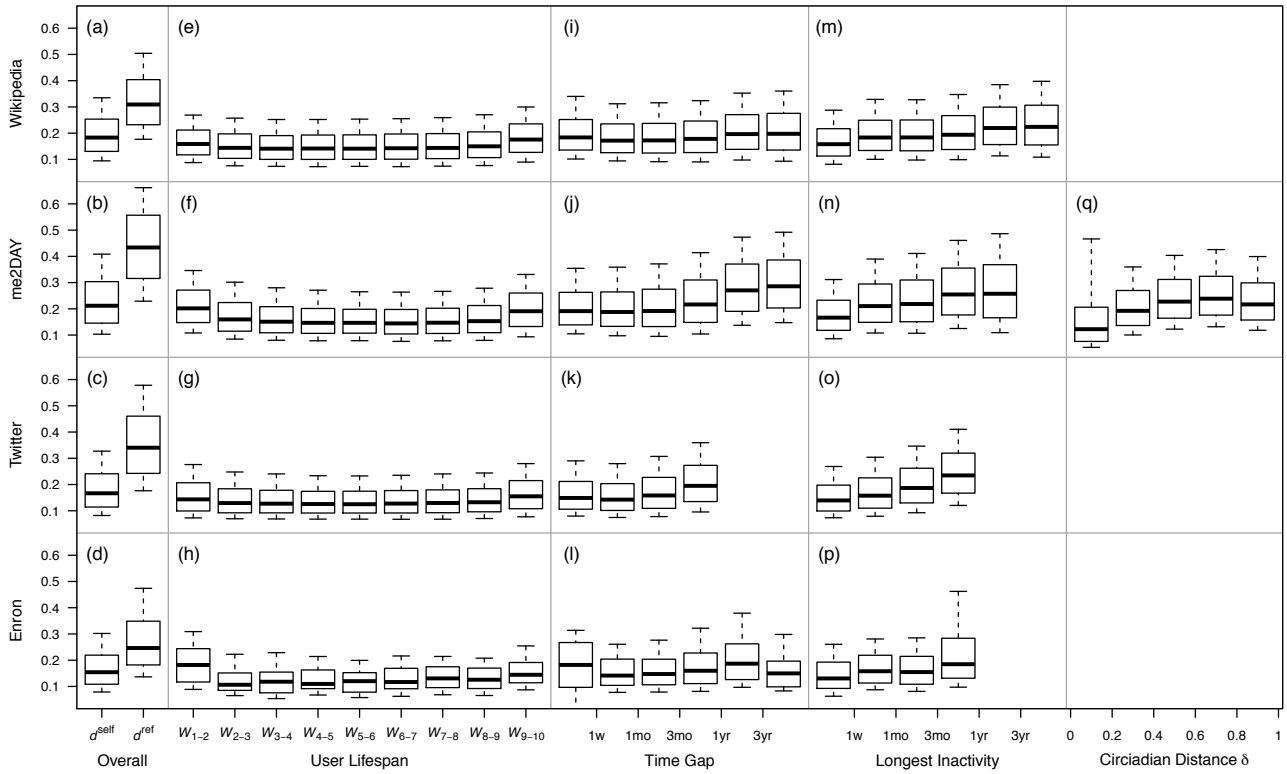



**Figure 2: Distribution of self-distances and reference distances, where the lower and upper bars denote 10 and 90 percentiles, respectively. (a-d) Distribution of self-distances and reference distances between every pair of windows for all users. (e-h) Distribution of self-distances between two consecutive windows $W_i$ and $W_{i+1}$. (i-l) Distribution of self-distances grouped by the time gap between two windows $W_i$ and $W_j$, calculated as $\sum_{i<k<j} \sum_{\tau \in W_k} \tau$. (m-p) Distribution of self-distances grouped by longest inactivity period between two windows $W_i$ and $W_j$, calculated as $\max_{\tau \in W_k, i<k<j} \tau$. (q) Distribution of self-distances of interval patterns grouped by distance between circadian rhythms.**

Twitter, and Enron email, the self-distance between consecutive windows over user lifespans remain very stable, except at the beginning and end of the service membership. That is, the boxplots between $W_1$ and $W_2$ and between $W_9$ and $W_{10}$ have higher median and wider percentile values than the rest. These phases roughly correspond to a new user developing a unique activity pattern and an experienced user losing interest, respectively.

## 4.2 Over time

We have shown that a user's interval pattern stays persistent along the lifespan. In this section, we measure how *long* each user's interval pattern stays persistent, and whether the interval pattern persists if there is a long inactivity period between the two activity windows.

We first define the time gap between two activity windows $W_i$ and $W_j$ of a user as the time between the last post of $W_i$ and the first post of $W_j$. In other words, the time gap between the two activity windows is equal to the sum of all inter-event time $\tau$ caught between the two windows, $\sum_{i<k<j} \sum_{\tau \in W_k} \tau$. Next, we define the longest inactivity period as the maximum inter-event time caught between the two windows, $\max_{\tau \in W_k, i<k<j} \tau$. Finally, we flatten the user dimension and plot the self-distance distribution between all pairs of activity windows, with respect to the length of temporal gap and the longest inactivity period between them, respectively.

As shown in Figures 2(i-l), each individual's interval pattern is robust over years. Even the time gap between two windows exceeds 3 years, the self-distances are much smaller than the reference distances. Also, even after coming back from long-term inactivity, the self-distance is visibly less than the reference distances in Figures 2(m-p). Note that we limit the longitudinal analysis on Twitter only up to 1-year span due to the Twitter API restriction.

## 4.3 Through changing daily routine

Why is an individual's interval pattern persistent over time? One possible explanation is that each person's circadian rhythm influences behavior dynamics, resulting in a unique interval pattern for every individual. Recent research has derived and modeled inter-event time distributions of human activities as the consequence of their circadian rhythms [14, 26].

If a circadian rhythm affects online human behavior, fluctuation in an individual's circadian rhythm should result in fluctuations in interval patterns as well. In this section, we

examine if the change in a user's circadian rhythm is related to the change in the interval pattern.

We extract the circadian rhythms, or circadian distributions, of users through the similar approach used to extract interval patterns. We measure at what time on a 24-hour cycle a user engages in the online activity, and plot the probability distribution on the 24-hour time scale. We extracted the distributions using von Mises kernel density estimation for each window of every user. (See Appendix A.3 for detail.) Thus for each user, we have 10 circadian distributions corresponding to 10 windows.

Then, we calculate the distance between the circadian distributions, $\delta$, for every pair of activity windows of a user. We define this as *circadian distance*. We use Jensen–Shannon distance, as we did in computing self distances of interval patterns. Greater circadian distance indicates greater fluctuations in his circadian rhythm over time.

On Figure 2(q), we plot the distribution of self distance between all pairs of activity windows with respect to circadian distance of me2DAY users. We present the results for me2DAY dataset only, because me2DAY provides local time information for every post, while other platforms provide only server time. On the contrary to the expectation, we observe that the users' interval patterns are robust to severe changes in circadian rhythms. This indicates that users' online behavior remains persistent even when individuals' biological clocks go out of sync due to shifts in sleep schedules or rearrangement of working hours.

# 5. CHANGE IN INTERVAL PATTERNS

In the previous section, we have presented multiple evidences that suggest persistence in individuals' interval patterns. Overall, an individual's interval pattern is persistent, but at times, we observe distinct changes in the interval pattern. This raises the following question: when do these changes occur and why? In this section, using me2DAY data, we focus on the such changes in the interval patterns over time.

## 5.1 Clustering the Interval Patterns

Before examining the transitions in the interval patterns of users, we first group the interval patterns into clusters. We perform $k$-means clustering [23] with $k = 12$ over all empirical windows observed in me2DAY in the previous section. The algorithm identifies 12 most representative cluster centroids, and maps every interval pattern into one of the twelve clusters. With these mappings, changes in interval patterns are abstracted as transitions between clusters, which enhances the interpretability of the analysis results.

Figure 3 shows the twelve clusters from $k$-means clustering for our me2DAY data as an example. Clusters $C_1$ to $C_{12}$ are sorted based on the expected value of logarithmic inter-event time, $E[\log_2 \tau]$. In other words, clusters with smaller index values correspond to *burstier* interval patterns in which activities repeatedly take place within a short span of time. Likewise, clusters with bigger index values resemble users who updates posts less frequently.

## 5.2 Interpreting Inter-cluster Transitions

Now that we have mapped the interval patterns into clusters, we analyze the transition between the interval patterns in this section. The transition between the interval patterns $P_i$ and $P_{i+1}$ of two consecutive windows $W_i$ and $W_{i+1}$ is

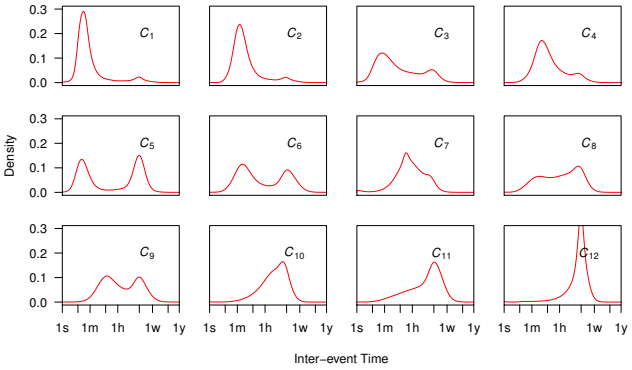



**Figure 3: The 12 centroids for clusters $C_1, ..., C_{12}$ from $k$-means clustering with $k = 12$ on all empirical interval patterns in me2DAY. The clusters are sorted by the expected value of logarithmized interval, $E[\log_2 \tau]$. Low index represents bursty posting behavior, and high index represents infrequent behavior.**

defined as follows:

$$(P_i \rightarrow P_{i+1}) \mapsto (src, dst, d, i) \quad (1)$$

where $src$ is the cluster that $P_i$ belongs to, $dst$ is the cluster that $P_{i+1}$ belongs to, $d$ is the distance between two interval patterns $P_i$ and $P_{i+1}$, and $i$ is the index of the prior window.

Figure 4(a) illustrates the extent to which the transition between two clusters occur. We measure the proportion of transitions from cluster $x$ to cluster $y$ over all transitions, formally denoted as $Pr(src = x, dst = y)$, and plot the probabilities. Note that the cells along the diagonal of the matrix, *i.e.* $C_i \rightarrow C_i$, signify those cases in which no transition occurs between two consecutive interval patterns. The figure serves as another evidence showing that activity patterns tend to be persistent in a large majority of cases, in accordance with our findings from the previous section.

A closer inspection of the transition between clusters showed a number of boundary data points moving in and out of neighboring clusters. To better focus on the more meaningful transitions, we only account for transitions with significant changes, by applying a self-distance threshold of 0.4, corresponding to the top 10% highest self-distances in Figure 2(b), to Figure 4(a). Each cell in Figure 4(b) represents $Pr(src = x, dst = y | d > 0.4)$, the ratio of $C_x \rightarrow C_y$ given consecutive window transitions with distance greater than 0.4.

We observe two notable characteristics from Figure 4(b). First, significant transitions mostly occur between $C_1/C_2$ and the other clusters. As established in the beginning of this section, $C_1$ and $C_2$ marks *bursty* updates with second- to minute-scale centroid modes. Second, the likelihood of bidirectional cluster transitions are symmetric, *i.e.* a person is just as likely to make transition $C_x \rightarrow C_y$ as he/she is likely to make transition $C_y \rightarrow C_x$. To recap, a transient, bursty behavior of a user triggers significant cluster transitions, and such transition is equally likely to be reversed. Manually inspecting several significant cluster transitions confirmed that the causes of such bursty behavior include posting along to a live event such as a sporting event or a television program, and writing down lyrics to a song.

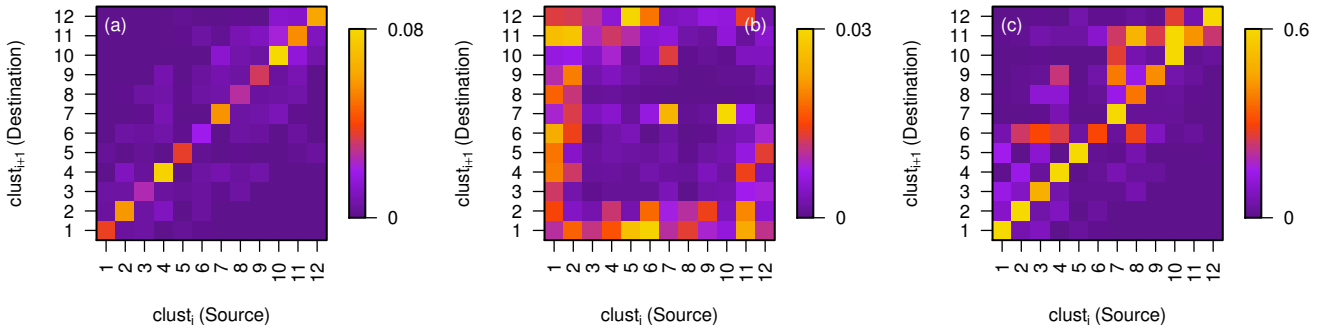



**Figure 4: The matrices showing transition probabilities. Each cell represents the transition probability from cluster $x$ to cluster $y$. (a) Joint probability matrix for transition between all pairs of clusters, $Pr(src = x, dst = y)$. (b) Joint probability matrix for transition between pairs of clusters whose distance is greater than 0.4, $Pr(src = x, dst = y|d > 0.4)$. (c) Transition probabilities only for the last pair of interval patterns, $Pr(dst = y|src = x, i = 9)$.**

The transition matrix is also useful in discovering accounts whose abandonment is imminent. To understand how users behave before leaving the service, we plot the transition probabilities just for the last pair of windows, $W_9$ and $W_{10}$. In other words, we compute $Pr(dst = y|src = x, i = 9)$ for all me2DAY users and display the results on Figure 4(c). Once again, probabilities are high along the diagonal, indicating that in most cases no transition occurs. Meanwhile, when transitions do occur, they are made from clusters of smaller index to greater index, as the upper-left half of the matrix is covered in lighter shades. It implies that the users' activity patterns tend to slow down prior to leaving the service.

## 6. APPLICATION: IDENTIFYING USERS

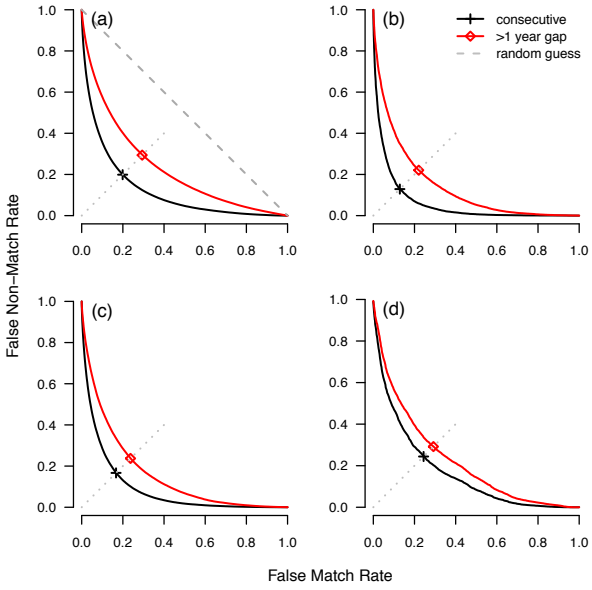
We have so far shown that individuals exhibit persistent and distinctive interval patterns of activities online. In this section, we exploit this insight to identify users only by observing the interval patterns. We formalize the user identification problem as the following. *Given two windows each containing 100 intervals, can we determine those are from the same user or not?*

For the task, we build a very simple threshold-based identifier. As input, the identifier takes two windows each containing 100 intervals. If the distance between two interval patterns is smaller than the threshold, it tells that those are from the same user. Otherwise, it tells those are from different users.

For the non-matching cases, we randomly select 10 window pairs for every pair of users. For the matching cases, (i) we randomly select 100 consecutive window pairs for each user, and (ii) we randomly select 100 window pairs with more than 1 year gap for each user. We split the matching cases into two groups to compare the identification performance over time gap.

To evaluate the performance, we use metrics typically used in biometric systems [22, 29, 35]. *False match rate* (FMR) measures the proportion of window pairs from a single user classified as from different users, and *false non-match rate* (FNMR) measures the proportion of window pairs from different users classified as from a single user. If we increase the threshold value, FNMR decreases but FMR increases. We refer the crossover point at which FMR equals FNMR as *equal error rate* (ERR).

Figure 5 shows the performance of the identifier with the detection error tradeoff (DET) curves. Considering that many implementations of behavioral biometric systems using key stroke dynamics, mouse dynamics, or gait recognitaion have ERR around 10% [8, 16, 29, 35], the user identification performance is remarkable. Only by observing the distance between interval pattern, we can match the same user's interval patterns with ERR around 20%, even there are long time gap exceeding 1 year.




| | Wikipedia | me2DAY | Twitter | Enron |
|---|---|---|---|---|
| Consecutive | 19.9% | 12.9% | 16.7% | 24.4% |
| >1 year gap | 29.4% | 22.1% | 23.7% | 29.2% |

**Figure 5: Detection error tradeoff curves for the user identification and the corresponding equal error rates. (a) Wikipedia. (b) me2DAY. (c) Twitter. (d) Enron.**

The performance of our prototype identifier suggest that interval pattern can serve as a behavioral feature in user identification. Only by comparing the interval patterns, we can determine that those are from the same individual or not, over time gap and across similar services.

Behavioral traces are hard to manipulate, while online profiles can be easily copied. Thus, we expect that the interval signature of online behavior can be used to detect online identity theft such as account hijacking or doppelgänger attack [5, 9].

# 7. RELATED WORK

The availability of time-stamped behavioral data has enabled researchers to study temporal human dynamics—often captured by inter-event time distributions, that we call *interval pattern* here. Aggregated inter-arrival time distributions matter greatly in system performance and service management. More specifically, inter-arrival time distributions are used to characterize web services [1, 2, 42] or categorize user sessions [10, 20].

Search for universality in natural phenomena including human behavior has been the edict of scientists. Statistical physicists have strived to mine a universal temporal pattern across all individuals, and explicate or model the behavior [3, 13, 17, 25, 26, 36, 41]. However, their observations sometimes compete with each other [4, 13, 33], suggesting that it is difficult to generalize interval patterns of individual human actions with a single formula.

In light of this, we focus on the persistence and distinctiveness of individuals' interval patterns rather than on finding a universal pattern. A previous study reported relative stability of temporal email patterns [24]. Focusing more on the longitudinal persistence of interval patterns in individual human behavior, our work scrupulously confirmed multiple datasets of various activities, spanning several years up to a decade and a half of time.

# 8. CONCLUSION

Characterizing a person, in a way, is discovering the individual's personal characteristics that are *invariant over time* and *different from others*. Behavioral signatures, as reflections of such persistent personal characteristics, have been the key component for observing and defining personality [27, 28, 32].

In this paper, we discover that each individual has a persistent *interval signature* that stays resilient under changing circumstances. We confirm this finding across different types of online behavior—article editing, microblog posting, and email. Our work demonstrates that the interval signature qualifies as a form of personal characteristics.

Based on our finding, we have applied the interval patterns in practical use of identifying users. Our prototype implementation demonstrates that the interval pattern is a good behavior feature distinguishing one from the others. In addition, our finding opens up many more interesting follow-up research questions: For a group of people having similar interval signatures, what do they have in common? What can be inferred about users by analyzing their interval patterns? How can we interpret the shape of an interval signature in terms of personality or other dimensions of personal characteristics? We believe that the notion of interval pattern introduced in this work is a valuable new measure in personal identification and categorization of online behaviors.

# 9. ACKNOWLEDGMENTS

We thank Yunkyu Sohn, Jahyun Goo, JinYeong Bak, and Hosung Park for useful discussion; and Jaimie Y. Park and Junhyun Shim for helping manuscript preparation. This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2012M3C4A7033342).

# 10. REFERENCES

[1] E. Adar, J. Teevan, and S. T. Dumais. Large scale analysis of web revisitation patterns. In *Annual SIGCHI Conference: Human Factors in Computing Systems*, CHI '08, pages 1197–1206, Apr. 2008.

[2] E. Adar, J. Teevan, and S. T. Dumais. Resonance on the web: Web dynamics and revisitation patterns. In *Annual SIGCHI Conference: Human Factors in Computing Systems*, CHI '09, pages 1381–1390, Apr. 2009.

[3] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, May 2005.

[4] A.-L. Barabási, K.-I. Goh, and A. Vázquez. Reply to comment on "The origin of bursts and heavy tails in human dynamics". *arXiv:physics/0511186*, Nov. 2005.

[5] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks. In *International World Wide Web Conference*, WWW '09, pages 551–560, Apr. 2009.

[6] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts. No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. In *International World Wide Web Conference*, WWW '13, pages 307–318, May 2013.

[7] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860, July 2003.

[8] D. Gafurov. *Performance and Security Analysis of Gait-based User Authentication*. PhD thesis, University of Oslo, May 2008.

[9] O. Goga, G. Venkatadri, and K. P. Gummadi. The Doppelgänger Bot Attack: Exploring Identity Impersonation in Online Social Networks. In *ACM Internet Measurement Conference*, IMC '15, pages 141–153, Oct. 2015.

[10] A. Halfaker, O. Keyes, D. Kluver, J. Thebault-Spieker, T. Nguyen, K. Shores, A. Uduwage, and M. Warncke-Wang. User Session Identification Based on Strong Regularities in Inter-activity Time. In *International World Wide Web Conference*, WWW '15, pages 410–418, May 2015.

[11] N. Hughes. 10 Celebs that Actually Run Their Own Twitter Account. The Huffington Post, Jan. 2014.

[12] J. Jeong, J.-H. Kang, and S. Moon. A Possible Explanation of SNS Dissolution from User Influx. In *Korea Computer Congress*, KCC '16, pages 1461–1463, June 2016.

[13] Z.-Q. Jiang, W.-J. Xie, M.-X. Li, B. Podobnik, W.-X. Zhou, and H. E. Stanley. Calling patterns in human communication dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 110(5):1600–1605, Jan. 2013.

[14] H.-H. Jo, M. Karsai, J. Kertész, and K. Kaski. Circadian pattern and burstiness in mobile phone communication. *New Journal of Physics*, 14(1):013055–18, Jan. 2012.

[15] M. C. Jones, J. S. Marron, and S. J. Sheather. A Brief Survey of Bandwidth Selection for Density Estimation. *Journal of the American Statistical Association*, 91(433):401–8, Mar. 1996.

[16] Z. Jorgensen and T. Yu. On Mouse Dynamics as a Behavioral Biometric for Authentication. In *ACM Asia*

*Conference on Computer and Communications Security*, AsiaCCS '11, pages 476–482, Mar. 2011.

[17] M. Karsai, K. Kaski, A.-L. Barabási, and J. Kertész. Universal features of correlated bursty behaviour. *Scientific Reports*, 2:397, May 2012.

[18] B. Klimt and Y. Yang. The Enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, ECML '04, pages 217–226, Sept. 2004.

[19] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, Mar. 1951.

[20] R. Kumar and A. Tomkins. A Characterization of Online Browsing Behavior. In *International World Wide Web Conference*, WWW '10, pages 561–570, Apr. 2010.

[21] J. Lin. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, Jan. 1991.

[22] S. Liu and M. Silverman. A practical guide to biometric security technology. *IT Professional*, 3(1):27–32, 2001.

[23] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[24] R. D. Malmgren, J. M. Hofman, L. A. N. Amaral, and D. J. Watts. Characterizing Individual Communication Patterns. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 607–616, June 2009.

[25] R. D. Malmgren, D. B. Stouffer, A. S. L. O. Campanharo, and L. A. N. Amaral. On Universality in Human Correspondence Activity. *Science*, 325(5948):1696–1700, Sept. 2009.

[26] R. D. Malmgren, D. B. Stouffer, A. E. Motter, and L. A. N. Amaral. A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences of the United States of America*, 105(47):18153–18158, Nov. 2008.

[27] W. Mischel. Toward an Integrative Science of the Person. *Annual Review of Psychology*, 55(1):1–22, Feb. 2004.

[28] W. Mischel, Y. Shoda, and R. Mendoza-Denton. Situation-behavior profiles as a locus of consistency in personality. *Current Directions in Psychological Science*, 11(2):50–54, 2002.

[29] A. Peacock, X. Ke, and M. Wilkerson. Typing Patterns: A Key to User Identification. *IEEE Security & Pravicy*, 2(5):40–47, 2004.

[30] J. Saramäki, E. A. Leicht, E. López, S. G. B. Roberts, F. Reed-Tsochas, and R. I. M. Dunbar. Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences of the United States of America*, 111(3):942–947, Jan. 2014.

[31] S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society Series B (Methodological)*, 53(3):683–690, 1991.

[32] Y. Shoda, W. Mischel, and J. C. Wright. Intraindividual Stability in the Organization and Patterning of Behavior: Incorporating Psychological Situations Into the Idiographic Analysis of Personality. *Journal of Personality and Social Psychology*, 67(4):674–687, 1994.

[33] D. Stouffer, R. Malmgren, and L. Amaral. Comment on "The origin of bursts and heavy tails in human dynamics". *arXiv:physics/0510216*, 2005.

[34] C. C. Taylor. Automatic bandwidth selection for circular density estimation. *Computational Statistics & Data Analysis*, 52(7):3493–3500, 2008.

[35] J. A. Unar, W. C. Seng, and A. Abbasi. A review of biometric technology along with trends and prospects. *Pattern Recognition*, 47(8):2673–2688, Aug. 2014.

[36] P. O. S. Vaz de Melo, C. Faloutsos, R. Assunção, and A. A. F. Loureiro. The self-feeding process: A unifying model for communication dynamics in the web. In *International World Wide Web Conference*, WWW '13, pages 1319–1330, May 2013.

[37] Wikipedia. Blocked users — Wikipedia, the free encyclopedia, 2016.

[38] Wikipedia. Wikipedia:List of bots by number of edits — Wikipedia, the free encyclopedia, 2016.

[39] Wikipedia. Wikipedia:Sock puppetry — Wikipedia, the free encyclopedia, 2016.

[40] Wikipedia. Wikipedia:Username policy — Wikipedia, the free encyclopedia, 2016.

[41] Y. Wu, C. Zhou, J. Xiao, J. Kurths, and H. J. Schellnhuber. Evidence for a bimodal distribution in human communication. *Proceedings of the National Academy of Sciences of the United States of America*, 107(44):18803–18808, Nov. 2010.

[42] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 177–186, Feb. 2011.

# APPENDIX

# A. METHODS

## A.1 Estimating Interval Patterns

Our goal is to estimate the probability density function of the inter-event interval $\tau$, given consecutive intervals in an activity window $W = (\tau_1, \tau_2, ..., \tau_n)$. We estimate the probability density function of the logarithimized inter-event time, $\log_2 \tau$, by a Gaussian kernel density estimator,

$$P = \hat{p}\left(\log_2 \tau\right) = \frac{1}{nh} \sum_{i=1}^{n} \Phi\left(\frac{\log \tau - \log \tau_i}{h}\right) \tag{2}$$

where $\Phi$ is the standard normal density function, and $h$ is the smoothing parameter called bandwidth. For the bandwidth, we basically follow Sheather and Jones' bandwidth selector as it well fits bimodal distributions [15, 31]. In addition, we set the upper and lower bounds on bandwidth at 1.5 and 0.5 to prevent over- and under-smoothing. We also tried other constant bandwidth of $1/2$, $1/\sqrt{2}$, and Epanechnikov kernel instead of Gaussian, but in all cases, the result was almost identical.

## A.2 Comparing Two Probability Densities

The Jensen–Shannon divergence (JSD) [21] is a popular measure of the difference between two probability density functions. It is a symmetrized and smoothed version of Kullback–Leibler divergence (KLD) [19]. JSD is defined by

$$\mathrm{JSD}(P||Q) = \frac{1}{2}\mathrm{KLD}(P||M) + \frac{1}{2}\mathrm{KLD}(Q||M) \tag{3}$$

where $M = (P + Q)/2$, and KLD is calculated as

$$\mathrm{KLD}(P||Q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx \tag{4}$$

Instead of directly using JSD, we use the square root of it $\sqrt{\mathrm{JSD}}$—called *Jensen–Shannon distance*—since it satisfies *metric* properties [7].

## A.3 Estimating Circadian Rhythms

To estimate circadian rhythms—*i.e.* what time on a 24-hour cycle a user engages in the activity—of online human behavior, we use the von Mises kernel density estimation [34]. For a given set of timestamps of posts in an activity window $W = (t_1, t_2, ..., t_n)$, the circadian rhythm $C$ is defined as

$$C = \hat{p_d}(t) = \frac{1}{n(24)I_0(\nu)} \sum_{i=1}^{n} \exp\{\nu \cos\left(\frac{2\pi(t - t_i)}{24}\right)\} \tag{5}$$

where $I_r(\nu)$ is the modified Bessel function of order $r$ and the concentration parameter $\nu$ is the inverse of the smoothing parameter $h$. Large values of $\nu$ lead to highly variable estimations whereas small values provided over-smoothed circular densities. For the concentration parameter, we use $\nu = 48$.

# B. EXAMINING WINDOW SIZE EFFECT

The self-distances and reference distances are key metrics in quantifying persistence and distinctiveness of interval patterns. However, the window size is directly related to the resolution of the interval pattern. If the window size is too small, the window might be too limited in time. If the window size is too large, the number of users who have more activities than the window size is decreases and use of the reference distance is limited to those few. Then, what size window is better in capturing both of persistence and diversity? In Figure 6, we vary the window size and plot the distribution of all self-distances and reference distances for all users.

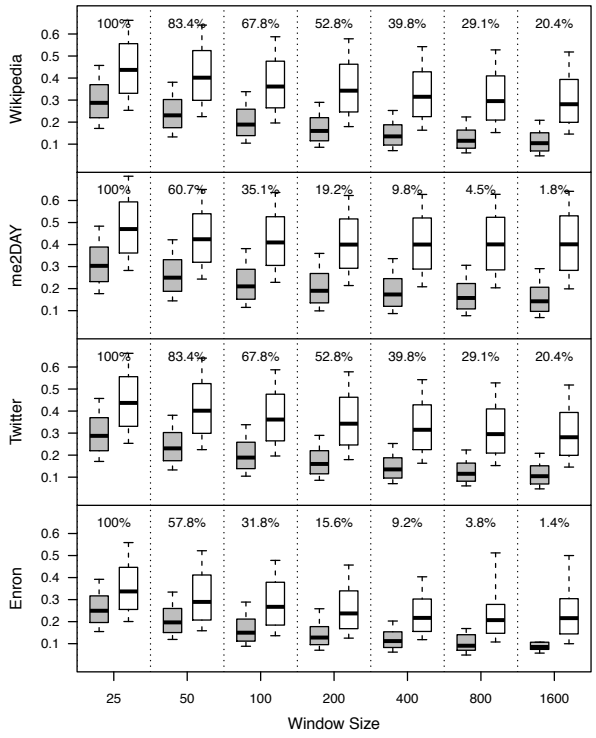



**Figure 6: The distributions of self-distance (gray) and reference distance (white) compared at different window sizes, where the lower and upper bars denote 10 and 90 percentiles, respectively. The numbers at the top are the proportion of users who punched more than $n$ actions.**

We use box-plots of 10 and 90 percentiles to summarize the distributions. The $x$-axis represents the window size from 25 to 1600. For each window size $n$, we select all users with more than $2n$ actions in each platform. For each of these users, we calculate 100 self-distances by randomly choosing non-overlapping window pairs, each of the same size $n$. The gray boxes represent the distributions of self-distances. As the window size increases, we observe that the self-distances decrease significantly.

The white boxes represent the distribution of reference distances. For each window size $n$, we select all users with more than $n$ actions in each dataset. Then, for each pair of users, we extract 100 pairs of windows randomly selected from each, and calculate the reference distances. The number at the top of a pair of gray and white boxes indicates the proportion of users with more than $n$ actions on each platform. As the window size increases, the cross-individual comparison is limited to a small number of highly active users. Accordingly, the reference distances decrease but the decrease in the range of percentiles is not as significant as in gray boxes.

Moreover, the overlap between two distributions become smaller as the window size increases. The result implies that each individual has unique interval signature, and we can estimate them more accurately as we observe more intervals.