RESEARCH ARTICLE

A Multiscale Survival Process for Modeling Human Activity Patterns

Tianyang Zhang[1], Peng Cui[1]*, Chaoming Song[2]*, Wenwu Zhu[1], Shiqiang Yang[1]

1 Department of Computer Science and Technology, Tsinghua University, Beijing, China, **2** Department of Physics, University of Miami, Miami, Florida, United States of America

* cuip@tsinghua.edu.cn (PC); csong@physics.miami.edu (CS)

Abstract

Human activity plays a central role in understanding large-scale social dynamics. It is well documented that individual activity pattern follows bursty dynamics characterized by heavy-tailed interevent time distributions. Here we study a large-scale online chatting dataset consisting of 5,549,570 users, finding that individual activity pattern varies with timescales whereas existing models only approximate empirical observations within a limited timescale. We propose a novel approach that models the intensity rate of an individual triggering an activity. We demonstrate that the model precisely captures corresponding human dynamics across multiple timescales over five orders of magnitudes. Our model also allows extracting the population heterogeneity of activity patterns, characterized by a set of individual-specific ingredients. Integrating our approach with social interactions leads to a wide range of implications.

Introduction

Human activity pattern is one of the central building blocks of modeling and understanding social dynamics such as information spreading [1–3], social-tie and group formations [4–7], social cooperations and competitions [8, 9]. While a wide range of social interaction models exist, they mostly assume that the communications among individuals are largely random, following a Poisson process. Yet, recent researches on human dynamics have demonstrated extensive evidence [10–12] that the interevent time (time between consecutive messages) and the response time (time between a message was received and the reply was sent) τ are heavy-tailed distributed, in contrast to prediction of the uncorrelated Poisson process where the interevent time distribution $P(\tau)$ follows an exponential form. This indicates that the vast majority of responses were sent within a very short time frame known as bursts. In some cases, however, the response stalls for a long time, as predicted by the long waiting times at the tail of the distribution. Specific examples range from communications [13, 14], entertainment [15, 16] and work patterns [17–19] to neural activities [20, 21], implying that there exists intrinsic complexity at each individual level even without involving social interactions. In other words, in social systems the "propagator" (interevent time distribution) $P(\tau)$ of a free "particle" (person) is fundamentally distinct from these in the physical science that are purely random (Guassian or

identifiable information according to their privacy policy. Users must agree with the privacy policy before using the Tencent QQ services. –Furthermore please clarify what restrictions on data availability exist as indicated in the submission form of this manuscript As we have signed agreement with Tencent company and promise that the data is only used for academic research, we can't upload it in a public repository. However, we are willing to share the data if any researcher contact us and sign an agreement with us about the usage of data under the terms and conditions of Tencent company.

Competing Interests: The authors have declared that no competing interests exist.

exponential). In contrast, the intrinsic complexity of human behaviors such as long-memory effect are encoded and translated into the non-trivial form of $P(\tau)$, which significantly impacts on social dynamics at a macroscopic level. For instance, it has been subsequently shown [22] that the non-Poisson nature of the contact dynamics fundamentally alters spreading processes on networks, resulting in notably larger decay times than predicted by Poisson processes.

To capture underlying complexity in human activity patterns, various models have been developed during the past decade. Overall these models fall into two major approaches. The first approach mainly focuses on the microscopic foundation of human dynamics and tends to model how individuals make actions or responses. For instance, Barabási proposed a simple queuing model that allows to capture some essential ingredients of bursty dynamics [23–26]. These models mostly predict a power law $P(\tau)$ with universal exponent 1 and 1.5 for fixed and variable queues, respectively. The second approach, however, tends to model $P(\tau)$ directly without involving microscopic information at the individual level. Candidate models include Weibull distribution, log-normal distribution and Pareto distribution that follows power law for all τ greater than a threshold x_m [27–29]. These models have more practical flexibility compared to queuing models yet lack microscopic understandings of human dynamics.

To demonstrate the challenges of modeling human activity pattern precisely, we plot $P(\tau)$ in Fig 1a for one online chatting user where τ is the interevent time of two consecutive messages (see Datasets section for details). The interevent time τ ranges over five orders of magnitudes and there is no simple distribution being able to approximate $P(\tau)$ across the entire time scale. For instance, a power law only captures an intermediate time regime $\tau \in (10^2, 10^4)$ approximately, and one has to crop data to perform power law fitting [30]. To quantify the effects of cropping, we fit $P(\tau)$ with the Pareto distribution for the same user which discards interevent time τ smaller than a parameter x_m. Fig 1b shows the fraction of cropped data versus fitting goodness, showing that the power law fits well only after cropping 30% of the data. More severely, when apply the Pareto distribution to the population, less than 10% of the users pass statistical test even when croping 70% of the data. Fig 1 reveals the fact that there exist different time scales where individuals' actions and responses have remarkably distinct patterns. This finding calls the needs for generic yet accurate models that enable to capture and quantify human dynamics across full-time scale. In this work, we report such a generic framework through a multiscale survival process.

Materials and Methods

Datasets

Our data-driven approach relies on accessibility of the following large-scale datasets of human activity pattern.

Online chatting: This data is collected from Tencent QQ, an MSN-style instant message platform in China, covering over 600 million users. We collect the log of users' group chatting behavior that contains 50,000 online groups with 5,549,570 group members and the message log of each group during a 2 month period. The data records a unique user ID for every individual user and all timestamps when the user posts messages, which allows to construct each user's activity pattern. All data is collected by Tencent QQ for academic research according to its terms and conditions. Furthermore, the data is fully anonymous and contains no identifiable information.

We also apply our method to the following two well-studied small-scale datasets for comparison and cross-validation:

Letter correspondence: The dataset is collected in the same way as in [24], and we adopt the similar analysis strategy. We collect 28511 letters of Einstein and 6944 letters of Freud, apart
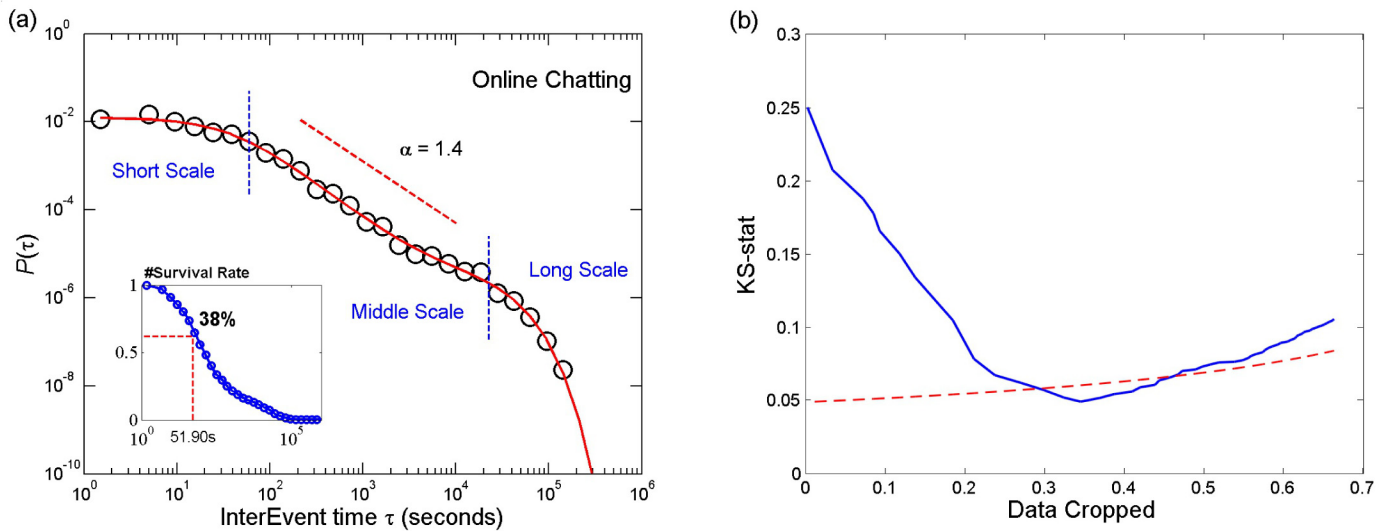
Fig 1. (a) The probability density function of the time interval between any user's two consecutive messages in online chatting. $P(\tau)$ is well approximated by a power-law $\tau^{-\alpha}$ with the exponent $\alpha = -1.4$ for the intermediate time regime $\tau \in [10^2, 10^4]$ whereas for both shorter and longer time scales the power-law characteristic does not work. Left bottom inset: survival rate of the time interval. (b) When fitting with Pareto distribution, we discard the data with interevent time less than x_m. We tune the parameter x_m and find the best fitting. It shows the relation between the fraction of cropped data less than x_m and the goodness of fit, which is measured by the KS statistic. The red line shows the threshold statistic to pass the KS test with the significance level of 5%, which gradually increases due to the decreased amount of data.

doi:10.1371/journal.pone.0151473.g001

from the ones missing date or sender/receiver. We calculate the response time of the letters and analyze the distribution of it.

Emails: The dataset contains 3188 users and 129 135 records of sending and receiving Emails during a 3 month period in a university environment [10]. We analyze the distribution of time intervals between an individual sending two consecutive e-mails as in [12, 24].

Model

The fact that the interevent time distribution $P(\tau)$ behaves differently at different time scales indicates that the intensity rate λ for an individual to make an action or response varies with the time τ. Therefore instead of modeling $P(\tau)$ directly, we are aiming to model the intensity function $\lambda(\tau)$ that encodes all essential microscopic details of the underlying stochastic process. Survival analysis [31–33] allows to connect λ to $P(\tau)$ through following rate equation

$$\frac{\partial S(\tau|\theta)}{\partial \tau} = -\lambda(\tau|\theta)S(\tau|\theta), \tag{1}$$

where the survival function $S(\tau|\theta)$ is the probability of a waiting time longer than τ, and $\{\theta\}$ are individual specific parameters. Solving Eq (1) leads to

$$S(\tau|\theta) = \exp\left(-\int_0^\tau \lambda(t|\theta)dt\right). \tag{2}$$

Given the waiting time distribution $P(\tau|\theta)$, the survival function $S(\tau|\theta)$ simply corresponds to its complementary cumulative distribution

$$S(\tau|\theta) = \int_\tau^\infty P(t|\theta)dt. \tag{3}$$

Combining Eqs ([2](#)–[3](#)) allows to predict

$$P(\tau|\theta) = -\frac{\partial S(\tau|\theta)}{\partial \tau} = \lambda(\tau|\theta) \exp\left(-\int_0^\tau \lambda(x|\theta)dx\right). \tag{4}$$

Here the $P(\tau|\theta)$ is a generic form of several simple distributions commonly used to model human dynamics. A time-independent intensity rate λ corresponds to a homogeneous Poisson process, where [Eq 4](#) simply recovers the well-known exponential waiting time distribution. If λ varies with time τ, we are able to achieve various $P(\tau)$ forms. For instance, $\lambda(\tau) = \gamma/\tau$ leads to a power law waiting time distribution $P(\tau) \sim \tau^{-(1+\gamma)}$ whereas $\lambda(\tau) = \gamma/\tau^\alpha$ with an exponent $\alpha < 1$ recovers the Weibull distribution. Nevertheless, as we discussed above, these simple forms capture only limited temporal regimes of the empirically observed $P(\tau)$.

To incorporate different activity patterns across multiple time scales, we propose the following generic intensity function

$$\lambda(\tau|\theta) = \frac{\lambda_0}{(\tau/t_0)^\alpha + 1} + \lambda_\infty, \tag{5}$$

where $\theta = \{\lambda_0, t_0, \alpha, \lambda_\infty\}$ captures the following different aspects in human activity patterns:

$\boldsymbol{\lambda_0}$ determines the activity rate in a small time scale. If $\tau \ll t_0$, $(\tau/t_0)^\alpha \approx 0$ and $\lambda_0 \gg \lambda_\infty$, we find $\lambda(\tau) \approx \lambda_0$. The larger λ_0 is, the higher the probability that the user makes a quick response will be.

$\boldsymbol{t_0}$ determines the critical time scale where a highly heterogeneous activity pattern starts to emerge. This phenomenon is well-known as the burstiness of human dynamics [[34](#), [35](#)]

$\boldsymbol{\alpha} > 0$ controls the degree of the heterogeneity of burst regime. The larger α value leads to more heterogeneous activities raised from underlying human dynamics.

$\boldsymbol{\lambda_\infty}$ determines the activity rate in a large time scale, e.g. $\lim_{\tau \to \infty} \lambda(\tau) = \lambda_\infty$. $\lambda_\infty \ll \lambda_0$ so that λ_∞ has little influence until τ is big enough. The exponential tail is greatly influenced by λ_∞ and it also plays a leading role in the average time interval of human activities.

To learn our modelling parameters, we start from a set of empirical records of an individual's interevent times $T = \{\tau_1, \tau_2, \ldots, \tau_n\}$, and calculate the following likelihood function

$$L(\theta) = \prod_{i=1}^n P(\tau_i|\theta) = \prod_{i=1}^n S(\tau_i|\theta)\lambda(\tau_i|\theta). \tag{6}$$

The corresponding log-likelihood function reads

$$\ln L(\theta) = \sum_{i=1}^n \left\{ \ln \lambda(\tau_i|\theta) - \int_0^{\tau_i} \lambda(t|\theta)dt \right\}. \tag{7}$$

Maximizing [Eq (7)](#) regarding $\{\lambda_0, t_0, \lambda_\infty, \alpha\}$ leads to estimated modeling parameters of the interevent times T.

Results

[Fig 2](#) demonstrates $P(\tau)$ for four randomly selected individuals across our datasets. Despite notable diversity over different individuals and datasets, our model excellently agrees with the empirical data across a full-time regime over five orders of magnitudes. In contrast, a power law fitting is often limited within an intermediate time-scale over 1–2 orders of magnitudes. The sharp cutoff at large time scale reveals a clear exponential tail in a semi-log plot in line with the prediction of our model ([Fig 2](#) Inset). Note that a typical timescale of such cutoff (e.g. start around 6 hours for online chatting dataset) is much shorter than the duration of data
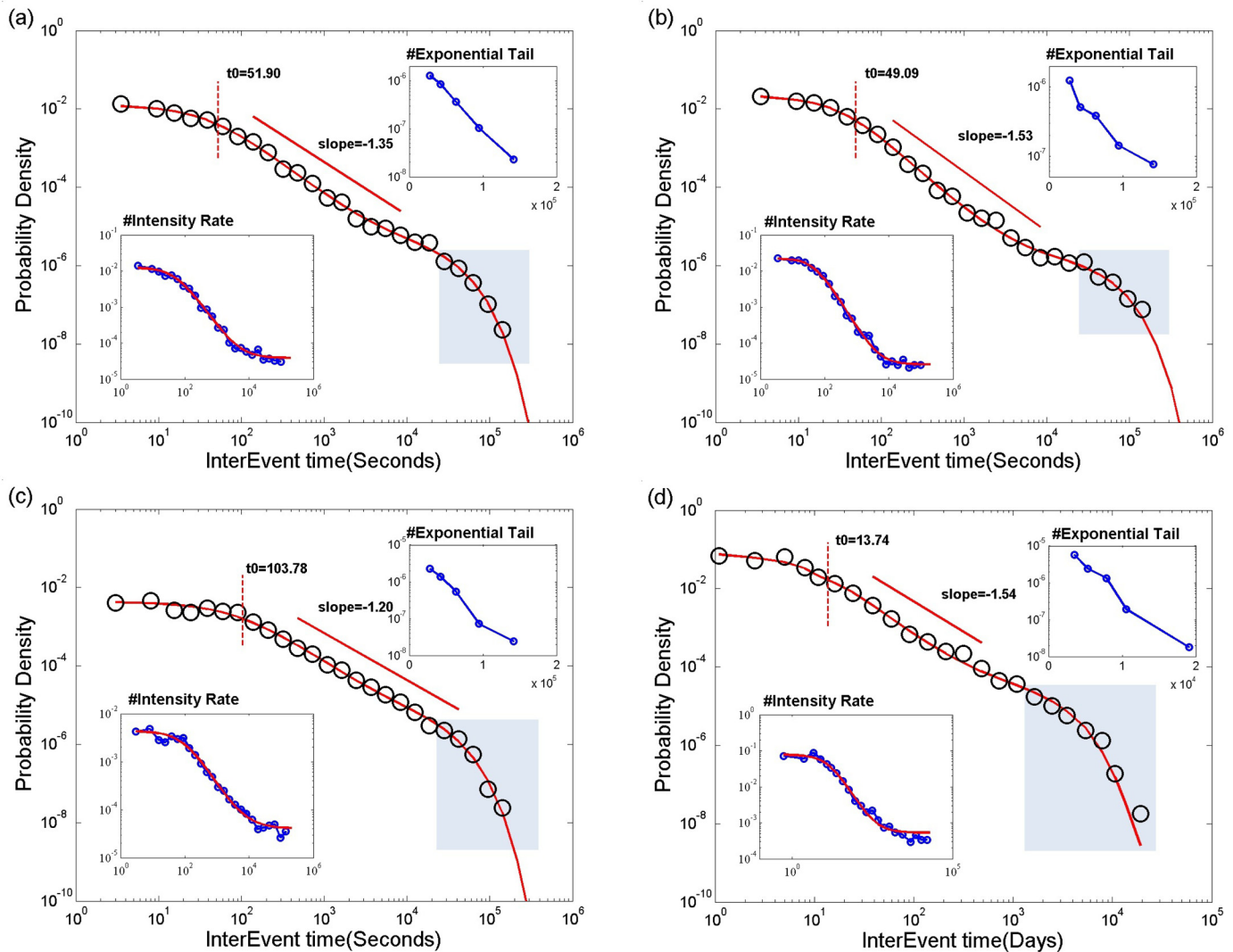
Fig 2. The interevent time distribution $P(\tau)$ for (a-b) two users from online chatting dataset, (c)one user for Email reply and (d) one individual for letter correspondence, respectively. The black circle represents the real data and the red curve is the model's prediction. Right top inset: $P(\tau)$ in a semi-log plot. Left bottom inset: intensity rate λ versus waiting time τ.

records(two months in our case), implying that it is rooted in the intrinsic ingredients of human activity instead of a finite-size effect. We also find that the time-dependent intensity rates $\lambda(\tau)$ monotonically decease with τ that are very well captured by our model Eq (5).

To evaluate the performance of our model, we apply our methodology to 26,648 active users (with more than 100 records during two month periods) from the online chatting dataset and apply several standard statistical tests including Kolmogorov-Smirnov test (KS test), chi-square test and Cramér-von Mises test. We set the significance level to 5% to judge whether the fitting is good enough or not. In addition, we use the average statistic as metric, which represents the magnitude of error between real data and the fitting result of the interevent time distribution.

Table 1 shows the average statistic of KS-test, chi-square test and Cramér test and the pass rate with the significance level of 5% in these three statistical tests on online chatting dataset.

As the table shows, our model beats all the baselines significantly in all metrics. 70.6% of users pass the KS-test with the significance level of 5%. If we lower the significance level to 1%, the pass

Table 1. Goodness of fit for the online chatting dataset, measured by the average statistic and pass rate with the significance level of 5% of three statistical testing method. Our model performs best in all metrics.

	KS test		chi-square		Cramer	
	stat	**rate**	**stat**	**rate**	**stat**	**rate**
Poisson	0.713	0.0%	8445.5	0.0%	79.43	0.0%
Pareto	0.281	0.5%	612.91	0.0%	19.91	0.5%
Weibull	0.202	1.7%	402.89	1.9%	5.46	2.0%
Log-normal	0.145	8.1%	216.24	3.7%	2.16	8.0%
Model	**0.070**	**70.6%**	**67.99**	**51.3%**	**0.66**	**76.54%**

rate will increase to 81.8% and it is quite an inspiring result. As for the bad case, we found that data sparsity is a main reason and observing more data will make the model more precise.

The improvement compared to baselines is mainly attributed to the truth that our model captures multi-time-scale characteristics of human dynamics in a more detailed and comprehensive way. The Poisson process performs the worst among all models since it cannot capture the heavy-tail feature of human dynamics. While Pareto distribution, Weibull distribution and log-normal distribution can capture heavy-tail feature, it only works for the middle time scale. As a result, if we try to fit the whole time scales of human dynamics, such models embody obvious limitation and are significantly exceeded by our model in all metrics. The high pass rate of our model also indicates that it captures multiple time scale human behaviour for real application.

We also compare the goodness of fit under different time scales after cropping data. Fig 3 shows the result of average KS p-value after cropping the short interevent time data on online chatting dataset and email dataset. We can see that the goodness of fit (measured by average p-value of KS test) all increases when gradually cropping data and stays relatively steady after discarding about 30%—40% of the data. The Pareto distribution, Weibull distribution and log-normal distribution show different patterns, reflecting partial information of human dynamics.
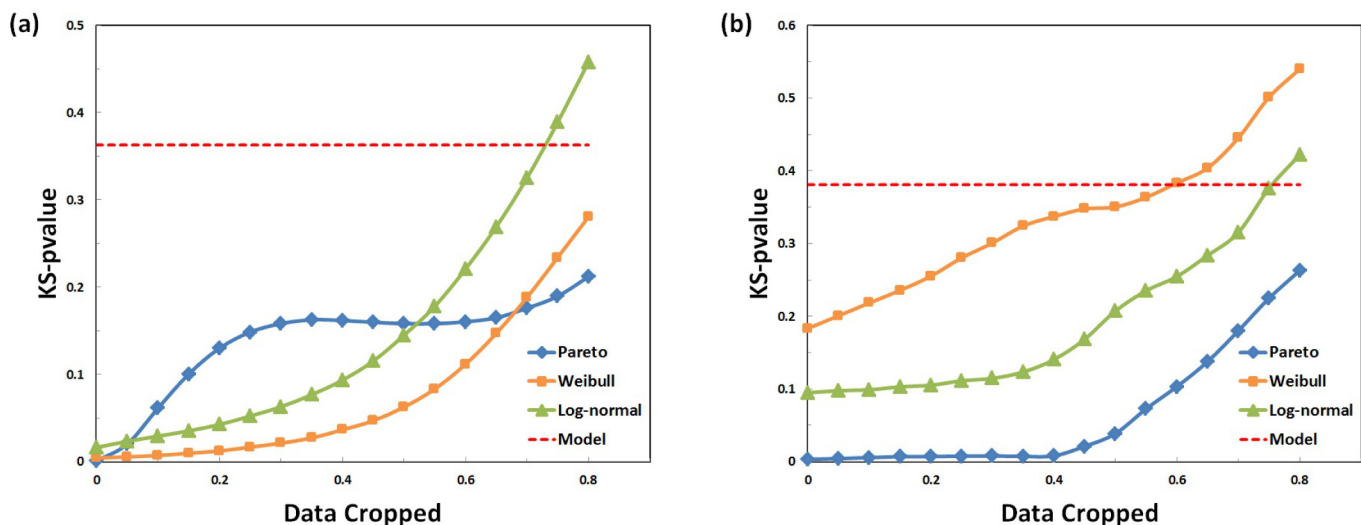

Fig 3. Average KS p-value after cropping, (a) online chatting dataset and (b) email activity dataset, respectively. The horizontal axis shows the fraction of data cropping. The vertical axis shows the average p-value in KS test, where higher value represents a better fitting. The blue, orange and green curves show the goodness of fit of Pareto distribution, Weibull distribution and log-normal distribution respectively. The red dotted line shows the p-value of our model without cropping data.
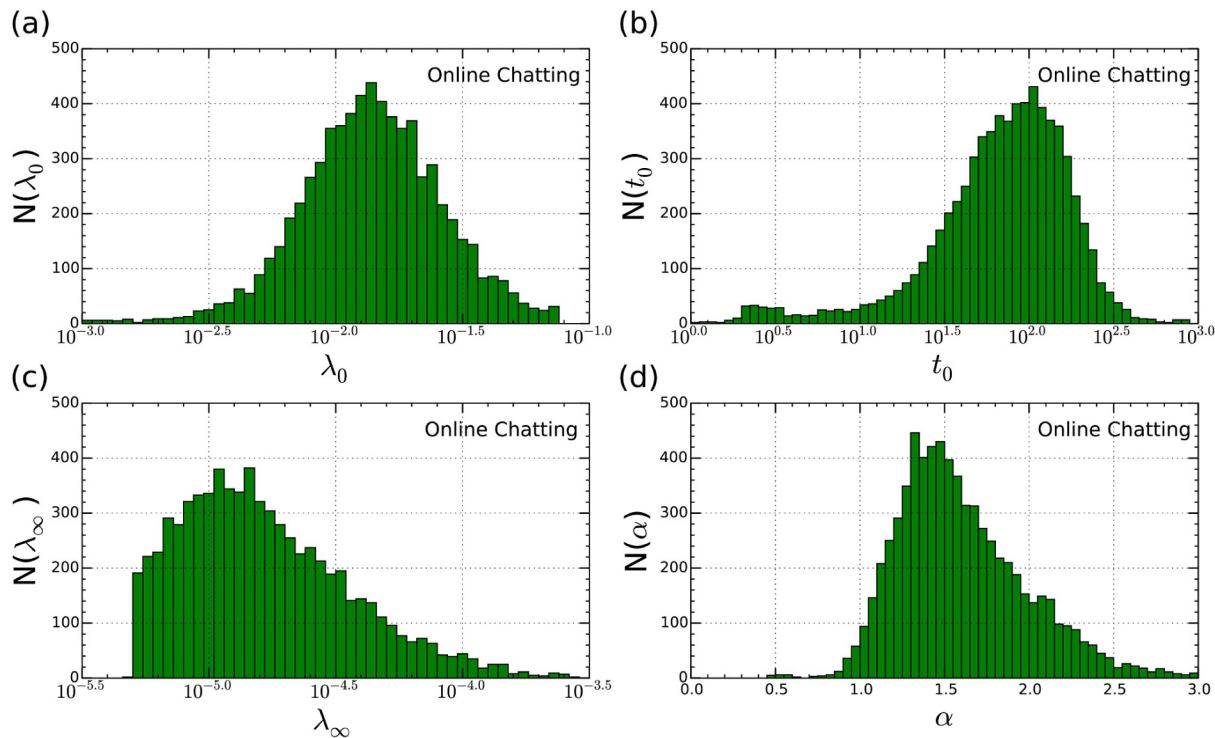
Fig 4. Distribution of parameter (a) λ_0, (b) t_0, (c) λ_∞ and (d) α for the online chatting dataset, respectively.

Pareto distribution performs worst for the full data yet improves significantly when cropping 30% of the data in short-time-scale. Log-normal distribution performs relatively stable and has advantage over Pareto distribution in shorter and longer time scales, implying that the early period and later period of human dynamics are non-power-law. The performance of our model (without cropping data) is represented by red dotted lines, showing remarkable improvements over all baselines.

Fig 4 plots the distributions of our modeling parameters across the online chatting dataset, finding that

λ_0 follows a log-normal distribution quite well. Since it mainly influences the percentage of short interevent intervals, we may infer that people's short time response patterns is stable around different users and the percentage of quick responses is also near Gaussian distribution;

t_0 follows a skewed log-normal distribution with 100 seconds as the modal number. As t_0 is the typical time scale of human activity between the quick response and bursty patterns, the plot indicates that for most people, the power-law distributed activity dynamics dominates after a relatively long period. We hypothesise that t_0 corresponds a time scale that an individual sticks to a certain topic whereas for the $\tau > t_0$, the user starts to lose his/her interests and switch to other topics. Further studies need to be performed to test this hypothesis.

λ_∞ has a strong correlation with the average time interval in human dynamics. It can be approximated by a log-normal distribution with a cutoff at small value, a fact due to the sampling bias of neglecting inactivity users.

α follows a skewed normal distribution with an average value close to 1.45, in line with the prediction of queue models with various queue size [24].

Discussion

To conclude, previous studies have shown that human dynamics is characterized by bursts of events and long periods of inactivity. Nevertheless, the nature of burst dynamics remains elusive. Our study of high resolution records of human interactive behavior provides an in-depth analysis of human dynamics, revealing non-Poisson temporal patterns that suggests a rethinking of mechanisms governing the human dynamics. Rather than focusing on limited time regimes, we find different patterns over multiple time-scales. We propose a generic model which not only captures the microscopic dynamics comprehensively but also predicts the interevent time distribution of each individual accurately. In this way, our model offers a generic modeling framework of the dynamics of human activities, potentially impacting a wide range of applications from marketing to education and politics.

Supporting Information

S1 File. A Multiscale Survival Process for Modeling Human Activity Patterns SI.
(PDF)

Acknowledgments

Author Contributions

Conceived and designed the experiments: PC CS TZ. Performed the experiments: TZ. Analyzed the data: TZ. Contributed reagents/materials/analysis tools: SY WZ. Wrote the paper: PC CS TZ. Gave advice on the manuscript: SY WZ.

References

1. Watts DJ. A simple model of global cascades on random networks. Proceedings of the National Academy of Sciences. 2002; 99(9):5766–5771. doi: 10.1073/pnas.082090499

2. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, et al. Identification of influential spreaders in complex networks. Nature Physics. 2010; 6(11):888–893. doi: 10.1038/nphys1746

3. Guille A, Hacid H, Favre C, Zighed DA. Information diffusion in online social networks: A survey. ACM SIGMOD Record. 2013; 42(2):17–28. doi: 10.1145/2503792.2503797

4. Wasserman S, Faust K. Social network analysis: Methods and applications. vol. 8. Cambridge university press; 1994.

5. Onnela JP, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, et al. Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences. 2007; 104(18): 7332–7336. doi: 10.1073/pnas.0610245104

6. Lin YR, Chi Y, Zhu S, Sundaram H, Tseng BL. Analyzing communities and their evolutions in dynamic social networks. ACM Transactions on Knowledge Discovery from Data (TKDD). 2009; 3(2):8. doi: 10.1145/1514888.1514891

7. Scott J. Social network analysis. Sage; 2012.

8. Johnson DW, Johnson RT. Cooperation and competition: Theory and research. Interaction Book Company; 1989.

9. Castellano C, Fortunato S, Loreto V. Statistical physics of social dynamics. Reviews of modern physics. 2009; 81(2):591. doi: 10.1103/RevModPhys.81.591

10. Eckmann JP, Moses E, Sergi D. Entropy of dialogues creates coherent structures in e-mail traffic. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101(40): 14333–14337. doi: 10.1073/pnas.0405728101 PMID: 15448210

11. Oliveira JG, Barabási AL. Human dynamics: Darwin and Einstein correspondence patterns. Nature. 2005; 437(7063):1251–1251. doi: 10.1038/4371251a PMID: 16251946

12. Barabasi AL. The origin of bursts and heavy tails in human dynamics. Nature. 2005; 435(7039):207–211. doi: 10.1038/nature03459 PMID: 15889093

13. Johansen A. Response time of internauts. Physica A: Statistical Mechanics and its Applications. 2001; 296(3):539–546. doi: 10.1016/S0378-4371(01)00202-3

14. Johansen A. Probing human response times. Physica A: Statistical Mechanics and its Applications. 2004; 338(1):286–291. doi: 10.1016/j.physa.2004.02.054

15. Huang J, Li C, Wang WQ, Shen HW, Li G, Cheng XQ. Temporal scaling in information propagation. Scientific reports. 2014; 4.

16. Yao Y, Tong H, Xu F, Lu J. Predicting long-term impact of CQA posts: a comprehensive viewpoint. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2014. p. 1496–1505.

17. Harder U, Paczuski M. Correlated dynamics in human printing behavior. Physica A: Statistical Mechanics and its Applications. 2006; 361(1):329–336. doi: 10.1016/j.physa.2005.06.079

18. Alfi V, Parisi G, Pietronero L. Conference registration: how people react to a deadline. Nature Physics. 2007; 3(11):746–746. doi: 10.1038/nphys761

19. Candia J, González MC, Wang P, Schoenharl T, Madey G, Barabási AL. Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical. 2008; 41(22):224015. doi: 10.1088/1751-8113/41/22/224015

20. Peyrache A, Khamassi M, Benchenane K, Wiener SI, Battaglia FP. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. Nature neuroscience. 2009; 12(7):919–926. doi: 10.1038/nn.2337 PMID: 19483687

21. Gallos LK, Sigman M, Makse HA. The conundrum of functional brain networks: small-world efficiency or fractal modularity. Frontiers in physiology. 2012; 3. doi: 10.3389/fphys.2012.00123 PMID: 22586406

22. Vazquez A, Racz B, Lukacs A, Barabasi AL. Impact of non-Poissonian activity patterns on spreading processes. Physical review letters. 2007; 98(15):158702. doi: 10.1103/PhysRevLett.98.158702 PMID: 17501392

23. Vazquez A. Exact results for the Barabási model of human dynamics. Physical review letters. 2005; 95(24):248701. doi: 10.1103/PhysRevLett.95.248701 PMID: 16384430

24. Vázquez A, Oliveira JG, Dezsö Z, Goh KI, Kondor I, Barabási AL. Modeling bursts and heavy tails in human dynamics. Physical Review E. 2006; 73(3):036127. doi: 10.1103/PhysRevE.73.036127

25. Blanchard P, Hongler MO. Modeling human activity in the spirit of barabasi's queueing systems. Physical Review E. 2007; 75(2):026102. doi: 10.1103/PhysRevE.75.026102

26. Vajna S, Tóth B, Kertész J. Modelling bursty time series. New Journal of Physics. 2013; 15(10): 103023. doi: 10.1088/1367-2630/15/10/103023

27. Pinder JE III, Wiener JG, Smith MH. The Weibull distribution: a new method of summarizing survivorship data. Ecology. 1978; p. 175–179.

28. Crow EL, Shimizu K. Lognormal distributions: Theory and applications. vol. 88. Dekker New York; 1988.

29. Arnold BC. Pareto distribution. Wiley Online Library; 1985.

30. Hosking JR, Wallis JR. Parameter and quantile estimation for the generalized Pareto distribution. Technometrics. 1987; 29(3):339–349. doi: 10.1080/00401706.1987.10488243

31. Kleinbaum DG, Klein M. Survival analysis. Springer; 1996.

32. Klein JP, Moeschberger ML. Survival analysis: techniques for censored and truncated data. Springer Science & Business Media; 2003.

33. Miller RG Jr. Survival analysis. vol. 66. John Wiley & Sons; 2011.

34. Kleinberg J. Bursty and hierarchical structure in streams. Data Mining and Knowledge Discovery. 2003; 7(4):373–397. doi: 10.1023/A:1024940629314

35. Leskovec J, McGlohon M, Faloutsos C, Glance NS, Hurst M. Patterns of Cascading behavior in large blog graphs. In: SDM. vol. 7. SIAM; 2007. p. 551–556.