Learning and Interpreting Complex Distributions in Empirical Data

Chengxi Zang*
CS Department, Tsinghua University
Beijing, China
chengxi.zang@gmail.com

Peng Cui
CS Department, Tsinghua University
Beijing, China
cuip@tsinghua.edu.cn

Wenwu Zhu
CS Department, Tsinghua University
Beijing, China
wwzhu@tsinghua.edu.cn

ABSTRACT

To fit empirical data distributions and then interpret them in a generative way is a common research paradigm to understand the structure and dynamics underlying the data in various disciplines. However, previous works mainly attempt to fit or interpret empirical data distributions in a case-by-case way. Faced with complex data distributions in the real world, can we fit and interpret them by a unified but parsimonious parametric model?

In this paper, we view the complex empirical data as being generated by a dynamic system which takes uniform randomness as input. By modeling the generative dynamics of data, we showcase a four-parameter dynamic model together with inference and simulation algorithms, which is able to fit and generate a family of distributions, ranging from Gaussian, Exponential, Power Law, Stretched Exponential (Weibull), to their complex variants with multi-scale complexities. Rather than a black box, our model can be interpreted by a unified differential equation, which captures the underlying generative dynamics. More powerful models can be constructed by our framework in a principled way. We validate our model by various synthetic datasets. We then apply our model to 16 real-world datasets from different disciplines. We show the systematic biases of fitting these datasets by the most widely used methods, and show the superiority of our model. In short, our model potentially provides a framework to fit complex distributions in empirical data, and more importantly, to understand their generative mechanisms.

CCS CONCEPTS

• **Mathematics of computing** → **Distribution functions**; *Stochastic processes*; • **Networks** → *Network dynamics*; Social media networks;

KEYWORDS

Complex Distribution; Heavy-tailed Distribution; Survival Analysis; Dynamic Model; Interpretability;

*Tsinghua National Laboratory for Information Science and Technology (TNList)

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08.
https://doi.org/10.1145/3219819.3220073

ACM Reference Format:
Chengxi Zang, Peng Cui, and Wenwu Zhu. 2018. Learning and Interpreting Complex Distributions in Empirical Data. In *Proceedings of The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3220073

1 INTRODUCTION

To fit empirical data distributions by parametric models and then interpret them in a generative way is a major scientific paradigm to understand the structure and generative dynamics underlying the data, which is widely used in various domains, ranging from biology [9], physics [1, 5], social science [12, 15], to computer science [8, 32], etc. For example, by investigating the power-law degree distribution of networks [2], the physicists found the network evolution dynamics in random networks. By examining the response time distribution of the correspondence patterns of Darwin and Einstein [18], or online collaborations [30], the social scientists tried to uncover the decision making dynamics of human behaviors. By fitting the data distributions with Gaussian mixture model [19], Bayesian methods [14], or even deep generative models [10], the computer scientists try to find clustered structures and generative dynamics of observed datasets. In short, this scientific paradigm is applied to a wide range of data science tasks.

However, previous works mainly try to fit or interpret the complex empirical data in a case-by-case way. For example, the Gaussian distribution is most widely used to fit narrow-tailed data distributions. A vast amount of literatures try to model the heavy-tailed data by power law distribution [5], Weibull distribution (or stretched exponential distribution) [11], and so on. Specific mixture models are also used to fit complex multi-scale distributions [23, 30, 32]. Deep generative networks like GAN exhibit limited power in fitting 1-D parametric distributions [24]. Thus, can we have a unified model to fit and to interpret various complex data distributions in the real world? Answering to this question is of vital importance.

In this paper, we try to fit complex distributions in empirical data by investigating their generative dynamics. The intuition behind our model is as follows: we view the empirical data with complex distributions as being generated from a dynamic system, which takes uniform randomness as input. Rather than modeling various complex distributions directly in a case-by-case way, we try to model their unified and possibly parsimonious generative dynamics, which generate all these complex distributions. One example is shown in Table 1: rather than fitting Gaussian, Exponential, Power-law, Stretched-Exponential, and their complex variants with complexities in multi-scale regimes in a case-by-case way, through a four-parameter dynamic model, we can capture all of them. More complex dynamic models can be constructed by our framework in

a principled way. Efficient inference method and simulation algorithm are provided. Furthermore, rather than a black-box model, we explain the generative dynamics of these complex distributions by a unified dynamic differential equation. As for the experiments, we analyze the properties of our model by various synthetic datasets, and further validate it by 16 empirical datasets from a wide range of disciplines. Our model fits all these complex empirical data accurately (Fig. 5). Our model potentially provides a framework to fit complex distributions observed in the real world, and more importantly, to understand their generative mechanisms. We summarize our contributions as follows:

- *Unification power:* We propose a general model to fit various complex distributions in empirical data, together with inference and simulation algorithms.
- *Parsimony:* Our model has only four parameters to capture multi-scale complexities in empirical distributions.
- *Interpretability:* Our model is interpreted by a unified generative dynamic equation. All the parameters have clear physical meanings.
- *Usefulness:* Our model fits various empirical datasets accurately, and can be generalized to more complex cases in a principled way.

The outline of the paper is: survey, model, mechanisms, experiments, discussion and conclusions. Reproducibility: The software and datasets are open-sourced at www.calvinzang.com.

2 RELATED WORK

We mainly review the related works in following two folds:

From simple to complex distributions in empirical data. Narrow-tailed distributions, like exponential distribution and Gaussian distribution, can be well captured by their mean and variance, which have well-studied underlying structure and dynamics. In contrast, heavy-tailed distributions, like power-law distribution, stretched-exponential distribution, log-normal distribution etc., exhibit larger even infinity variance, implying complex underlying structure and dynamics of data. Among the heavy-tailed distributions, the power-law distribution is the most famous one due to its scaling property [22] and generative mechanisms[2] . Extensive evidence and discussions of power-law distributions can be found in [13, 16]. Recently, more and more literatures found the distributions of empirical data are more complex than pure power law, ranging from human behavior data [26, 32], network data [3], to various datasets as showin in Fig. 5.

Fitting complex distributions. The maximum likelihood estimation, possibly with priors or regularizers, is used for fitting narrow-tailed distributions [14]. On another hand, deep generative networks like GAN are validated by fitting 1-D parametric distributions but exhibit large bias [24]. In contrast, fitting theories for complex distributions, say skewed or heavy-tailed ones [17], is not well established. Taking the most typical case - power law distribution - as an example, the visual inspection and least-square fitting are used to fit power law distribution at first. Later, the well-celebrated work [5] shows the bias of the least-square fitting method, and then propose a parametric method ($f(x) = \frac{\alpha_{PL}-1}{x_{min}}(\frac{x}{x_{min}})^{\alpha_{PL}}$, denoted as PL method) to fit power law distribution based on maximum likelihood principal. The PL method has been widely used to fit plausible power-law distributions by a large number of scientific papers. However, we find PL method shows large bias when detecting the power law signals in the real-world data as shown in Fig. 5. The origin of failures lies in the fact that PL method ignores the complexity in the real-world data [23, 26, 32]. How to fit and interpret various complex distributions in empirical datasets by a unified model is largely unknown.

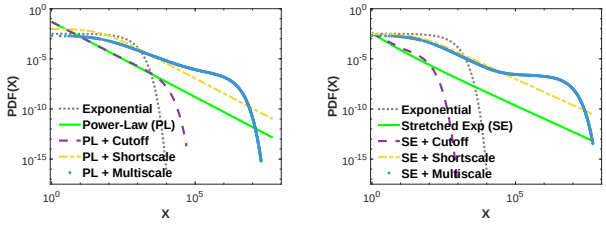
3 PROPOSED METHOD

3.1 Model Intuition

The intuition behind our model is as follows: we view the empirical data with complex distributions as being generated from a (nonlinear) dynamic system, which takes uniform randomness as input. Rather than modeling the complex outputs, namely various data distributions, of this dynamic system in a case-by-case way, we try to capture their unified generative dynamics. *In short, we try to model the simpler generative dynamics which generate complex phenomena.*

Our model is based on survival analysis [32], point process [26], and dynamic systems[25, 27, 31]. The probability density function $f_{\mathcal{X}}(x)$ of data $\mathcal{X} = (x_1, ..., x_{n-1}, x_n)$ can be modeled by the hazard function $\lambda(x) = \frac{f_{\mathcal{X}}(x)}{S_{\mathcal{X}}(x)}$, which describes the rate of the occurrence of the random variable $\mathcal{X} = x$ conditioned on $\mathcal{X} \geq x$, where $S_{\mathcal{X}}(x) = 1 - \int_{-\infty}^{x} f_{\mathcal{X}}(s)ds$. The $\Lambda(x) = \int_{-\infty}^{x} \lambda(s)ds$ represents the cumulative hazard rate. By modeling the hazard rate, we can get complex probability density function according to the relationship $f_{\mathcal{X}}(x) = \lambda(x)e^{-\int_{-\infty}^{x} \lambda(s)ds}$. We further build the link between hazard function $\lambda(x)$ and its corresponding dynamic system to interpret the generative mechanisms of the data distributions in the next section.

3.2 The Model



(a) $\theta = 1$ (b) $\theta \neq 1$

Figure 1: *Illustration of capabilities of our basic model.* Our model generates a family of distributions, including power law (PL), PL with cutoff, PL with short-scale complexity, PL with multi-scale complexity, exponential, stretched exponential (SE), SE with short-scale complexity, SE with multi-scale complexity, etc.

Here we propose the basic model, simple but versatile, leading to various distributions, as shown in Table. 1 and Fig. 1. The hazard function specifying the model is:

$$\lambda(x) = \beta + \alpha(x + \Delta)^{-\theta} \tag{1}$$

Table 1: Capability table. Our basic model encompasses all the following distributions. Illustrations are shown in Fig. 1.

Capability	Exponential	Power law	Power law + cutoff *	Power law + Shortscale	Power law + Multiscale
PDF ($\theta = 1$)	$\beta e^{-\beta x}$	$\alpha\Delta^{\alpha}x^{-(\alpha+1)}$	$\alpha\Delta^{\alpha}x^{-(\alpha+1)}e^{-\beta x}$	$\alpha\Delta^{\alpha}(x+\Delta)^{-(\alpha+1)}$	$(\beta+\frac{\alpha}{x+\Delta})(\frac{x}{\Delta}+1)^{-\alpha}e^{-\beta x}$
Hazard rate	β	$\frac{\alpha}{x}$	$\beta+\frac{\alpha}{x}$	$\frac{\alpha}{x+\Delta}$	$\beta+\frac{\alpha}{x+\Delta}$
Our model	✓	✓	✓	✓	✓
Capability	**Exponential**	**Stretched exponential** **	**Stretched exponential + Cutoff** *	**Stretched exponential + Shortscale**	**Stretched exponential +Multiscale**
PDF ($\theta \neq 1$)	$\alpha e^{-\alpha x}$	$\alpha x^{-\theta}e^{-\frac{\alpha}{1-\theta}x^{1-\theta}}$	$\alpha x^{-\theta}e^{-\frac{\alpha}{1-\theta}x^{1-\theta}-\beta x}$	$\alpha(x+\Delta)^{-\theta}e^{-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta}-\Delta^{1-\theta}]}$	$[\beta+\alpha(x+\Delta)^{-\theta}]e^{-\beta x-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta}-\Delta^{1-\theta}]}$
Hazard rate	α	$\frac{\alpha}{x^{\theta}}$	$\beta+\frac{\alpha}{x^{\theta}}$	$\frac{\alpha}{(x+\Delta)^{\theta}}$	$\beta+\frac{\alpha}{(x+\Delta)^{\theta}}$
Our model	✓	✓	✓	✓	✓

* For the Power law distritbuion with cutoff case and Stretched exponential distribution with cutoff case, the probability density functions of which are derived approximately by the harzard rates. Refer to the Model Section.

** Special case is approximately Normal Distribution when $\theta = -1$. Refer to the Model Section.

3.2.1 Critical Case: Power-Law-Based Distributions. When $\theta = 1$, $\lambda(x|\theta = 1)$ generates a family of distributions based on power-law distribution.

Lemma 3.1. *The $\lambda(x|\theta = 1)$ generates a family of distributions, from exponential distribution, power-law distribution, power-law distribution with exponential cut-off, to the complex multi-scale distributions.*

Proof. The probability density function of the random variable x is:

$$\begin{aligned} f(x|\theta=1) &= \lambda(x|\theta=1)e^{-\int_0^x \lambda(s|\theta=1)ds} \\ &= (\beta+\frac{\alpha}{x+\Delta})e^{-\beta x-\alpha\ln(\frac{x}{\Delta}+1)} \\ &= \beta e^{-\beta x}(\frac{x}{\Delta}+1)^{-\alpha}+\frac{\alpha}{\Delta}(\frac{x}{\Delta}+1)^{-(\alpha+1)}e^{-\beta x} \end{aligned} \tag{2}$$

Exponential distribution. When $\alpha = 0$, regardless of other three parameters, the $\lambda(x|\alpha = 0)$ generates exponential distribution with the probability density function $f(x|\alpha = 0) = \beta e^{-\beta x}$, as shown in Fig. 1a gray curve.

Power-law distribution. When $\beta = 0$ and $\Delta \ll x$, $f(x|\theta = 1, \beta = 0) = \alpha\Delta^{\alpha}(x+\Delta)^{-(\alpha+1)} \propto x^{-(\alpha+1)}$. Another meaning of Δ is the minimal value, say x_0, which x can take: $f(x|\theta = 1, \Delta = 0) = \frac{\alpha}{x}e^{\int_{x_0}^{x}\frac{\alpha}{s}ds} = \alpha {x_0}^{\alpha}x^{-(\alpha+1)}$.

Power-law distribution with cutoff. When $\beta \gg 0$ and $\Delta \ll x \ll \frac{\alpha}{\beta}-\Delta$, $f(x|\theta = 1) = (\beta+\frac{\alpha}{x+\Delta})(\frac{x}{\Delta}+1)^{-\alpha}e^{-\beta x} \approx \alpha\Delta^{\alpha}x^{-(\alpha+1)}e^{-\beta x}$.

Complex multi-scale distribution. When $\beta \to 0$, the complex multi-scale distribution features constant short-scale, power-law middle-scale and exponential long-scale. When $x \to 0$, $f(x|\theta = 1) \to \frac{\alpha}{\Delta}$. In the short-scale regime $x \in (0, \Delta]$, $f(x|\theta = 1) \approx \frac{\alpha}{\Delta}(\frac{x}{\Delta}+1)^{-(\alpha+1)}$, which decays slowly to the power-law middle-scale regime. When $\beta e^{-\beta x}(\frac{x}{\Delta}+1)^{-\alpha} \gg \frac{\alpha}{\Delta}(\frac{x}{\Delta}+1)^{-(\alpha+1)}e^{-\beta x}$, namely $x \gg \frac{\alpha}{\beta}-\Delta$, $f(x|\theta = 1) = \beta e^{-\beta x}(\frac{x}{\Delta}+1)^{-\alpha}+\frac{\alpha}{\Delta}(\frac{x}{\Delta}+1)^{-(\alpha+1)}e^{-\beta x} \approx \beta e^{-\beta x}(\frac{x}{\Delta}+1)^{-\alpha}$, which is the exponential long-scale regime. When $\Delta \ll x \ll \frac{\alpha}{\beta}-\Delta$, $f(x|\theta = 1) \approx \alpha\Delta^{\alpha}(x+\Delta)^{-(\alpha+1)} \propto x^{-(\alpha+1)}$, which is the power-law middle-scale regime. □

3.2.2 General Case: Stretched-Exponential-based Distribution. When $\theta \neq 1$, $\lambda(x|\theta \neq 1)$ generates a family of distributions based on stretched exponential distribution.

Lemma 3.2. *The $\lambda(x|\theta \neq 1)$ generates a family distributions, from exponential distribution, stretched exponential (Weibull) distribution, stretched exponential distribution with exponential cut-off, to the complex multi-scale distributions.*

Proof. Similar to the justification above, the probability density function of random variable x is:

$$\begin{aligned} f(x|\theta\neq1) &= \lambda(x|\theta\neq1)e^{-\int_0^x \lambda(s|\theta\neq1)ds} \\ &= [\beta+\alpha(x+\Delta)^{-\theta}]e^{-\beta x-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta}-\Delta^{1-\theta}]} \\ &= \beta e^{-\beta x}e^{-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta}-\Delta^{1-\theta}]} \\ &+ \alpha(x+\Delta)^{-\theta}e^{-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta}-\Delta^{1-\theta}]}e^{-\beta x} \end{aligned} \tag{3}$$

Exponential distribution. When $\beta = \theta = 0$, $f(x|\beta = 0, \theta = 0) = \alpha e^{-\alpha x}$.

Stretched exponential (Weibull) distribution. When $\beta = 0$ and $\Delta = 0$, the $\lambda(x|\theta \neq 1)$ leads to

$$f(x|\theta\neq1) = \alpha x^{-\theta}e^{-\frac{\alpha}{1-\theta}x^{1-\theta}} \tag{4}$$

The cumulative density function is $F(x|\theta \neq 1) = 1 - e^{-\frac{\alpha}{1-\theta}x^{1-\theta}}$, which is the stretched exponential distribution. Some special cases: exponential distribution $\alpha e^{-\alpha x}$ when $\theta = 0$, approximately *Normal* distribution $\frac{\alpha}{x^2}e^{-\frac{\alpha x^2}{2}}$ when $\theta = -1$.

Stretched exponential distribution with exponential cut-off. When $\beta \gg 0$ and $\Delta \ll x \ll (\frac{\alpha}{\beta})^{\frac{1}{\theta}}-\Delta$, $f(x|\theta \neq 1) = [\beta+\alpha(x+\Delta)^{-\theta}]e^{-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta}-\Delta^{1-\theta}]}e^{-\beta x} \approx \alpha x^{-\theta}e^{-\frac{\alpha}{1-\theta}x^{1-\theta}}e^{-\beta x}$.

Complex multi-scale distribution. When $\theta \neq 1$, the complex multi-scale distribution is based on stretched exponential distribution. When $\beta \to 0$, the complex multi-scale distribution features constant short-scale, stretched exponential middle-scale and exponential long-scale. When $x = 0$, $f(x|\theta \neq 1) \approx \frac{\alpha}{\Delta^{\theta}}$. In the short-scale regime $x \in (0, \Delta]$, $f(x|\theta \neq 1) \approx \alpha(x+\Delta)^{-\theta}e^{-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta}-\Delta^{1-\theta}]}$, which decays slowly to the stretched exponential middle-scale regime. When $\beta \gg \alpha(x+\Delta)^{-\theta}$, i.e., $x \gg (\frac{\alpha}{\beta})^{\frac{1}{\theta}}-\Delta$, $f(x|\theta \neq 1) \approx \beta e^{-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta}-\Delta^{1-\theta}]}e^{-\beta x}$, which is the exponential long-scale regime. When $\Delta \ll x \ll (\frac{\alpha}{\beta})^{\frac{1}{\theta}}-\Delta$, $f(x|\theta \neq 1) \approx \alpha(x+\Delta)^{-\theta}e^{-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta}-\Delta^{1-\theta}]}$, which is the stretched exponential middle-scale regime.

□

We illustrate above justifications in Fig. 1. We find complex distributions based on power law distribution (Fig. 1a) and stretched exponential distribution (Fig. 1b) are generated by our simple hazard function, possibly taking on complexities in multi-scale regimes.

3.3 Parameter Inference

The parameters of our model can be learned by the maximum likelihood estimation (MLE) framework. The log-likelihood function of observing a set of data $\{x_1, ..., x_{n-1}, x_n\}$ is given by:

$$\ln L(x_1, \ldots, x_n) = \ln \prod_{i=1}^{n} \lambda(x_i) e^{-\Lambda(x_i)} = \sum_{i=1}^{n} \ln \lambda(x_i) - \sum_{i=1}^{n} \Lambda(x_i) \tag{5}$$

According to the value of θ, $\Lambda(x)$ takes on different forms. When $\theta \neq 1$, the log-likelihood function is:

$$\ln L(x_1, \ldots, x_n | \theta \neq 1) = \sum_{i=1}^{n} \ln[\beta + \alpha(x_i + \Delta)^{-\theta}] - \beta \sum_{i=1}^{n} x_i - \frac{\alpha}{1-\theta} \sum_{i=1}^{n} [(x_i + \Delta)^{1-\theta} - \Delta^{1-\theta}] \tag{6}$$

, while when $\theta = 1$, the log-likelihood function is:

$$\ln L(x_1, \ldots, x_n | \theta = 1) = \sum_{i=1}^{n} \ln[\beta + \alpha(x_i + \Delta)^{-1}] - \beta \sum_{i=1}^{n} x_i - \alpha \sum_{i=1}^{n} \ln(\frac{x_i}{\Delta} + 1) \tag{7}$$

Maximizing Eq. 5 or 6 regarding $\{\beta, \alpha, \Delta, \theta\}$, subject to $\{\beta, \alpha, \theta \geq 0; \Delta > 0\}$ leads to estimated modeling parameters. However, due to the clear physical meaning of the parameters, prior knowledge can be applied to the initialization. We show this point later.

Another good advantage of the model is that all the parameters have closed-form gradients. The gradients for the $\theta \neq 1$ case, namely stretched exponential based model, are:

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} \frac{1}{A(i)} - \sum_{i=1}^{n} x_i \tag{8}$$

$$\frac{\partial \ln L}{\partial \alpha} = \sum_{i=1}^{n} \frac{(x_i + \Delta)^{-\theta}}{A(i)} - \frac{1}{1-\theta} \sum_{i=1}^{n} [(x_i + \Delta)^{1-\theta} - \Delta^{1-\theta}] \tag{9}$$

$$\frac{\partial \ln L}{\partial \Delta} = -\alpha\theta \sum_{i=1}^{n} \frac{(x_i + \Delta)^{-\theta-1}}{A(i)} - \alpha \sum_{i=1}^{n} [(x_i + \Delta)^{-\theta} - \Delta^{-\theta}] \tag{10}$$

$$\frac{\partial \ln L}{\partial \theta} = -\alpha \sum_{i=1}^{n} \frac{(x_i + \Delta)^{-\theta} \ln(x_i + \Delta)}{A(i)} - \alpha \sum_{i=1}^{n} \{\frac{[-(x_i + \Delta)^{1-\theta} \ln(x_i + \Delta) + \Delta^{1-\theta} \ln \Delta]}{1-\theta} + \frac{(x_i + \Delta)^{1-\theta} - \Delta^{1-\theta}}{(1-\theta)^2}\} \tag{11}$$

where $A(i) = \beta + \alpha(x_i + \Delta)^{-\theta}$.

When $\theta = 1$, the gradients for the power-law based model are:

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} \frac{1}{B(i)} - \sum_{i=1}^{n} x_i \tag{12}$$

$$\frac{\partial \ln L}{\partial \alpha} = \sum_{i=1}^{n} \frac{(x_i + \Delta)^{-1}}{B(i)} - \sum_{i=1}^{n} \ln(\frac{x_i}{\Delta} + 1) \tag{13}$$

$$\frac{\partial \ln L}{\partial \Delta} = -\alpha \sum_{i=1}^{n} \frac{(x_i + \Delta)^{-2}}{B(i)} + \alpha \sum_{i=1}^{n} \frac{x_i}{x_i \Delta + \Delta^2} \tag{14}$$

where $B(i) = \beta + \frac{\alpha}{x_i + \Delta}$. We can solve the optimization problem by many gradient-based optimization algorithms. For example, we adopt the interior point algorithm [4], and for reproducibility, we open our code, see Section 7.

3.4 Generator

The simplest and the most elegant way of generating a random number x from the cumulative distribution function $F(x)$ is the inverse transformation method [7]. First we generate a random number u from standard uniform distribution $U(0, 1]$. By solving the equation $F(x) = u$ for x, x is the number which follows distribution $F(x)$. We extend this inverse transformation method to the hazard rate function by the fact that $F(x) = 1 - e^{-\Lambda(x)}$, where $\Lambda(x) = \int_{x_0}^{x} \lambda(s)ds$. Thus, $F(x) = u = 1 - e^{-\Lambda(x)}$, and we can get the wanted random number by solving $\Lambda(x) = -\ln(1-u)$ for x where u and $1-u$ make no difference when sampling from $U(0, 1]$. Due to the fact that $\Lambda(x)$ is monotone-increasing function thus with inverse function Λ^{-1} , we can get

$$x = \Lambda^{-1}(-\ln u) \tag{15}$$

Even if there is no closed-form inverse function Λ^{-1}, we can get it numerically by solving the equation $\ln u + \Lambda(x) = 0$ for x where u is generated from the uniform distribution $U(0, 1)$.

Algorithm 1: Generating random samples specified by the hazard rate Equation 1

Input : Hazard function of model $\lambda(x) = \beta + \alpha(x + \Delta)^{-\theta}$, total number N
Output: $\{x_1, ..., x_N\}$

```
Set current number of events n = 1;
while n ≤ N do
    Sample u ~ Uniform([0, 1]) ;
    Solve ln u + Λ(x) = 0 for x by Algorithm 2.;
    x_n = x;
end
```

Algorithm 2: Newton's iterative method

Input : Equation $\Phi(x) = \log u + \Lambda(x)$.
Output: x

```
Set ε = 10^{-8}, x = 0;
while |Φ(x)| ≤ ε do
    if θ == 1 then
        Φ(x) = ln u + βx + α ln(x/Δ + 1);
    else
        Φ(x) = ln u + βx + α/(1-θ) [(x + Δ)^{1-θ} - Δ^{1-θ}];
    end
    Φ'(x) = β + α(x + Δ)^{-θ} ;
    x = x - Φ(x)/Φ'(x) ;
end
```

4 PHYSICAL MECHANISMS

In this section, we give the underlying generative dynamics of our model, namely Equ. 1 together with various distributions as shown in Table 1. We view the complex data distributions as being generated by a (non-linear) dynamic system, which takes uniform randomness as input:

4.1 Uniform Inputs and Growth

In order to give a dynamic view of the data generative process, our first step is to figure out the inputs of the dynamic system by connecting the point process and survival analysis. The process by which we sample n numbers from the standard uniform distribution $U(0, 1]$ can be viewed as a stochastic point process. Given a Poisson process $N(t) = \{t_i | i = 1, ..., N(t) = n; t_i \leq t_2 \leq ... \leq t_n\}$, then t_i is

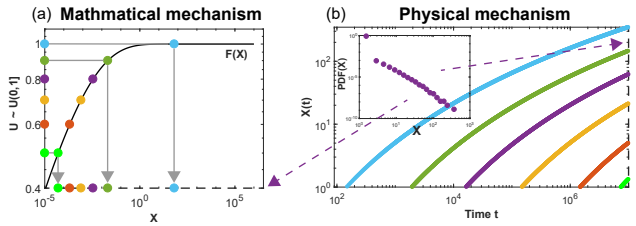


Figure 2: ***Illustration of (a) generator and (b) generative dynamics for generating stretched-exponential data as shown in the inset in (b). Generated samples by the generator are distributed in x-axis, while samples by physical process are distributed as a cross-sectional snapshot along x(t) axis.***

uniformly distributed on the interval $(0, t]$. If we normalize the t_i by t, then $u = \frac{t_i}{t}$ follows the standard uniform distribution $U(0, 1]$. We replace u in Equ. 15 by $\frac{t_i}{t}$, leading to growth dynamics of an agent i with a uniformly arriving time t_i in $(0, t]$:

$$x_i(t) = \Lambda^{-1}(-\ln u) = \Lambda^{-1}(-\ln(\frac{t_i}{t})) = \Lambda^{-1}(\ln(\frac{t}{t_i})). \tag{16}$$

For instance, let $\lambda(x) = \alpha x^{-\theta}$, which generates power-law distribution $f(x) = \alpha\Delta^{\alpha}x^{-(\alpha+1)}$ when $\theta = 1$, and stretched-exponential distribution $f(x) = \alpha x^{-\theta}e^{-\frac{\alpha}{1-\theta}x^{1-\theta}}$ when $\theta \neq 1$ (Detailed in Sec. 3). Thus, $\Lambda(x|\theta = 1) = \alpha\int_{\Delta}^{x}\frac{1}{s}ds = \alpha\ln(\frac{x}{\Delta})$, and $\Lambda(x|\theta \neq 1) = \alpha\int_{\Delta}^{x}\frac{1}{s^{\theta}}ds = \frac{\alpha(x^{1-\theta}-\Delta^{1-\theta})}{1-\theta}$. We can get their inverse function $\Lambda^{-1}(y|\theta = 1) = \Delta e^{\frac{y}{\alpha}}$, and $\Lambda^{-1}(y|\theta \neq 1) = (\frac{1-\theta}{\alpha}y + \Delta^{1-\theta})^{\frac{1}{1-\theta}}$.

By using $\Lambda^{-1}(y|\theta = 1) = \Delta e^{\frac{y}{\alpha}}$ to Equ. 16, we get growth curve over time t:

$$x_i(t) = \Lambda^{-1}(\ln(\frac{t}{t_i})|\theta = 1) = \Delta e^{\frac{\ln(\frac{t}{t_i})}{\alpha}} = \Delta(\frac{t}{t_i})^{\frac{1}{\alpha}}. \tag{17}$$

Similarly, by applying $\Lambda^{-1}(y|\theta \neq 1) = (\frac{1-\theta}{\alpha}y + \Delta^{1-\theta})^{\frac{1}{1-\theta}}$ to Eq. 16 when $\theta < 1$, we get growth curve:

$$x_i(t) = \Lambda^{-1}(\ln(\frac{t}{t_i})|\theta \neq 1) = (\frac{1-\theta}{\alpha}\ln(\frac{t}{t_i}) + \Delta^{1-\theta})^{\frac{1}{1-\theta}}. \tag{18}$$

4.2 Generative Dynamics: Base Model

Our second step is to reverse engineer generative dynamics by connecting the survival analysis and dynamic systems. We take derivative of Equ. 17 and Eq. 18 to time t, we get the generative dynamics of power-law distribution data and stretched-exponential distribution data as follows:

$$\frac{dx_i(t)}{dt} = \frac{d\Delta(\frac{t}{t_i})^{\frac{1}{\alpha}}}{dt} = \frac{\Delta}{t_i^{\frac{1}{\alpha}}}\frac{1}{\alpha}t^{\frac{1}{\alpha}-1} = \frac{x_i(t)}{\alpha t} \tag{19}$$

$$\frac{dx_i(t)}{dt} = \frac{d(\frac{1-\theta}{\alpha}\ln(\frac{t}{t_i}) + \Delta^{1-\theta})^{\frac{1}{1-\theta}}}{dt} = \frac{(\frac{1-\theta}{\alpha}\ln(\frac{t}{t_i}) + \Delta^{1-\theta})^{\frac{\theta}{1-\theta}}}{\alpha t} = \frac{x_i(t)^{\theta}}{\alpha t} \tag{20}$$

We find the linear preferential attachment by Eq. 19 to generate power-law distribution, and non-linear preferential attachment by Equ. 20 to generate stretched-exponential distribution, consistent with literatures on the scale-free observations in random networks [2].

4.3 Generative Dynamics: General Model

Here, we give the generative dynamics of our model. When $\lambda(x) = \beta + \alpha(x + \Delta)^{-\theta}$, $\Lambda(x) = \int_0^x \lambda(s)ds = \beta x + \frac{\alpha}{1-\theta}[(x + \Delta)^{1-\theta} - \Delta^{1-\theta}]$. By our construction, we get:

$$\Lambda(x_i(t)) = \beta x_i(t) + \frac{\alpha}{1-\theta}[(x_i(t) + \Delta)^{1-\theta} - \Delta^{1-\theta}] = \ln\frac{t}{t_i} \tag{21}$$

By taking derivative of above equation to time t, we get:

$$\frac{dx_i(t)}{dt} = \frac{(x_i(t) + \Delta)^{\theta}}{\beta(x_i(t) + \Delta)^{\theta}t + \alpha t} \tag{22}$$

Thus, from a dynamic view, the complex data which exhibit complex multi-scale distributions (captured by our model Equ. 1) are generated from a dynamic system with differential equation 22, including physical mechanisms: non-linear preferential attachment $(x_i(t)+\Delta)^{\theta}$, growth system αt, together with short-term complexity Δ and long-term complexity $\beta(x_i(t) + \Delta)^{\theta}t$. Our dynamics encompass Equ. 19 and 20 as special cases.

Thus, we describe the data generation process in random network scenario as follows:

- A new node i comes to the network following a Poisson process at $0 < t_i < t$ where t is the maximum observational time,
- and the degree of node i, denoted as $x_i(t)$, grows over time according to the differential equation 22 .

Then, the cross-sectional degree distributions of this network at time t follows $f_X(x) = \lambda(x)e^{-\int_{-\infty}^{x}\lambda(s)ds}$, where $\lambda(x) = \beta + \alpha(x + \Delta)^{-\theta}$ as shown in Equ. 1.

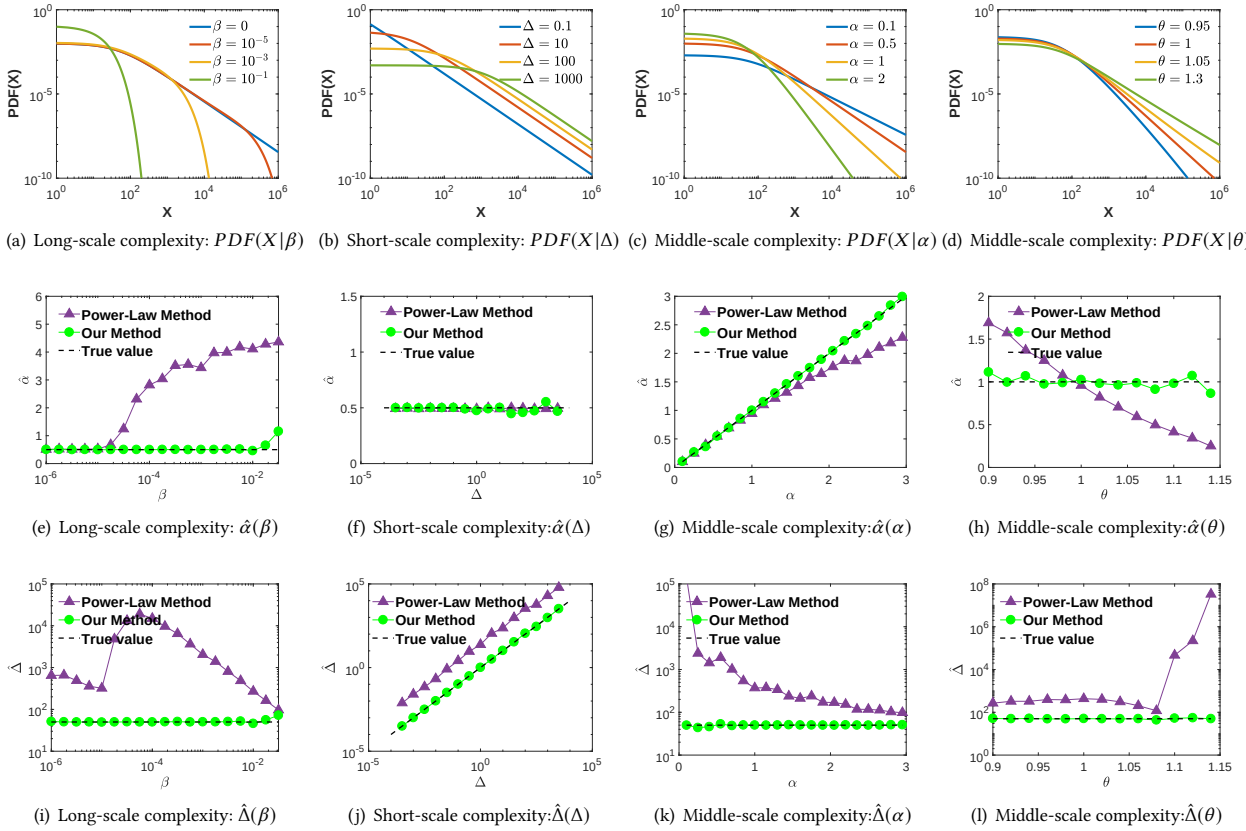
5 EXPERIMENTS

In this section, we evaluate our model on both synthetic and real-world datasets.

5.1 Synthetic data analysis

5.1.1 Ignoring the compleity leads to systematic bias. Actually, distributions which measure the quantities of the real-world data are much more complex than a pure power-law distribution. We show the evidence from the real-world datasets in next section. Here, we examine the possible biases which introduced by applying the well celebrated power-law fitting method (denoted as PL method) [5] to the complex distributions.

Long-scale complexity. We first investigate the long-scale complexity. The parameter β acts as the simplest form of modeling the long-scale complexity. By varying the β in $\lambda(x|\beta, \alpha = 0.5, \Delta = 50, \theta = 1)$, we get a series of distributions illustrated in Fig. 3a. The case when $\beta = 0$, as shown by the straight line part (with power-law exponent $1 + \alpha = 1.5$) of the blue curve in Fig. 3a, indicates the missing of long-scale complexity. As β increases, the long-scale complexity will move towards to the short-scale until the two parts overlap. Indeed, the characteristic scale of the long-scale regime is $\frac{\alpha}{\beta}$ and the characteristic scale of short-scale regime is Δ (Refer to Sec. 3). We avoid this overlap by limiting $\beta < 10^{-2}$. We generate 10^4 (a relatively large dataset to get reasonable fitting results and at the same time within the scalability limits of the baseline method. We will show the scalability of our model and the baseline later.) samples by the $\lambda(x|\beta, \alpha = 0.5, \Delta = 50, \theta = 1)$ with each specific β, and fit the α and Δ by the PL method and our method. Figures 3e



(a) Long-scale complexity: $PDF(X|\beta)$ (b) Short-scale complexity: $PDF(X|\Delta)$ (c) Middle-scale complexity: $PDF(X|\alpha)$ (d) Middle-scale complexity: $PDF(X|\theta)$

(e) Long-scale complexity: $\hat{\alpha}(\beta)$ (f) Short-scale complexity:$\hat{\alpha}(\Delta)$ (g) Middle-scale complexity:$\hat{\alpha}(\alpha)$ (h) Middle-scale complexity:$\hat{\alpha}(\theta)$

(i) Long-scale complexity: $\hat{\Delta}(\beta)$ (j) Short-scale complexity:$\hat{\Delta}(\Delta)$ (k) Middle-scale complexity:$\hat{\Delta}(\alpha)$ (l) Middle-scale complexity:$\hat{\Delta}(\theta)$

Figure 3: *Complexities in short-, long-, and middle- scale regimes introduce systematic biases to the previous method. Our method fits reality well.* The first row shows the distributions with different complexities in different regimes. Last two rows shown mean estimated values for the modeling parameters plotted as a function of the varying parameter which capture the complexity in corresponding regime. Figures in each column are of same settings. In all cases, the statistical errors are smaller than the data points. The true parameter values are shown in dashed lines. The pdf described in PL method is $f(x) = \frac{\alpha_{PL}-1}{x_{min}}(\frac{x}{x_{min}})^{\alpha_{PL}}$, where $\alpha = \alpha_{PL} - 1$

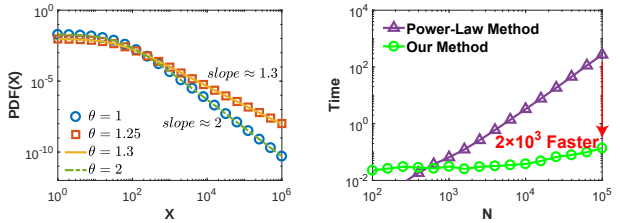
and i plot the mean estimated values of α (scaling exponent) and Δ as a function of β. We find the larger and larger discrepancy between the power-law scaling parameter α estimated by the PL method and the true value marked in dashed line as β increases as shown in Fig. 3e. In contrast, our model well gets the true scaling value. As for the estimation of short-scale Δ shown in Fig. 3i, the PL method seriously overestimates the true value, up to 450 times as much as true value when $\beta \approx 3*10^{-5}$. As β increases, the long-scale regime squeezes into the short-scale regime, and thus the estimated Δ by PL method decreases to the true value.

Short-scale complexity. We then consider the impact of short-scale complexity. The parameter Δ is the simplest form to capture the characteristic scale of the short-scale regime. Figure 3b plots $\lambda(x|\beta = 0, \alpha = 0.5, \Delta, \theta = 1)$ as Δ varies, which consists of short-scale regime followed by a power-law regime with same slope. The larger Δ is, the larger short scale is and thus a wider range of short-scale regime. Similarly, we generate 10^4 samples by the $\lambda(x|\beta = 0, \alpha = 0.5, \Delta, \theta = 1)$ with each specific Δ, and fit the α and Δ by the PL method and our method. Our model fits the real data quite well for both the parameters as shown in Figs. 3f and j. The PL method does well in fitting scaling exponent α for this special experimental setting, but the good result of α is at the cost of seriously overestimating Δ, indicating that the PL method discards the data samples in the short-scale regime which account for $\geq 80\%$ of the whole dataset.

Middle-scale complexity. Last but not least, we examine the middle-scale complexity. When $\theta = 1$, the middle-scale regime follows power law with scaling exponent $1 + \alpha$. By varying the α of the $\lambda(x|\beta = 0, \alpha, \Delta, \theta = 1)$ when control for other parameters, we get different scaling exponent of power law distribution at middle- and long- scale regimes. Figure 3c plots $\lambda(x|\beta = 0, \alpha, \Delta = 50, \theta = 1)$ as α varies, the larger the α, the steeper the curves. However, we find as α grows, the PL method underestimates the scaling parameter α, and the discrepancy becomes larger and larger, shown in Fig. 3g. Besides, the PL method seriously overestimates the short-scale parameter δ at the same time, up to 3 order of magnitudes, shown in Fig. 3k. In contrast, our model consistently reaches the true values well.

When $\theta \neq 1$, the middle-scale regime follows stretched-exponential law. Figure 3d plots $\lambda(x|\beta = 0, \alpha = 1, \Delta = 50, \theta)$ as θ varies. The

red curve with $\theta = 1$ is the power law distribution (at middle- and long- scale regimes) with slope of the pdf curve $\alpha + 1 = 2$, while other curves are stretched exponential distributions. We can not tell differences between these power law or stretched exponential curves by visual inspections. The power-law tools and the Taylor expansion can be used to analyze the complexities in this regime approximately in asymptotic analysis section later. Within the range of $(-\infty, 1 + \epsilon]$ where $\epsilon \to 0$ (refer to asymptotic analysis section), larger the θ, fatter the tail of the curves as shown in Fig. 3d. We find PL method overestimates the scaling parameter α when $\theta < 1$, and underestimates α when $\theta > 1$ shown in Fig. 3h. At the same time, PL method consistently overestimates the Δ as shown in Fig. 3l. In contrast, our model again gives much better estimations.



(a) β (b) β

Figure 4: (a) Asymptotic behaviors of the stretched-exponential distribution. (b) Scalability. Our model can be applied to large datasets while the power law method cannot.

5.1.2 Asymptotic behaviors. We analyze the asymptotic behaviors of our model. Taking the stretched exponential as an example, the asymptotic behaviors of which leads to the confusion with power law distribution and complex scaling law of the distribution. The pdf. of the stretched exponential distribution is $f(x|\theta \neq 1) = \alpha x^{-\theta} e^{-\frac{\alpha}{1-\theta} x^{1-\theta}}$ for $x > 0$. By applying the Taylor expansion when $x \to 0$ and $\theta < 1$, we get:

$$f(x \to 0|\theta < 1) = \frac{\alpha}{x^{\theta}}[1 - \frac{\alpha}{1-\theta} x^{1-\theta} + O(x^{2-2\theta})] \tag{23}$$

Thus, $f(x \to 0|\theta < 1) \approx \frac{\alpha}{x^{\theta}}$, which, however, can be easily lost by the short-scale regime when Δ is large. When $x \to \infty$, $f(x \to \infty|\theta < 1) = \alpha x^{-\theta} e^{-\frac{\alpha}{1-\theta} x^{1-\theta}}$, which decays faster than power law decay $\frac{\alpha}{x^{\theta}}$ but slower than exponential decay $e^{-\frac{\alpha}{1-\theta} x^1}$.

When $\theta > 1$ and $x \to \infty$, we get the Taylor expansion:

$$f(x \to \infty|\theta > 1) = \frac{\alpha}{x^{\theta}}[1 + \frac{\alpha}{(\theta-1)x^{\theta-1}} + O(\frac{1}{x^{2\theta-2}})] \approx \alpha \frac{1}{x^{\theta}} + \frac{\alpha^2}{\theta - 1} \frac{1}{x^{2\theta-1}} \tag{24}$$

, which approximates the power-law distribution with scaling exponent being governed by scaling parameter θ and $2\theta - 1$ alternatively. When $\theta = 1 + \epsilon$ where $\epsilon \to 0^+$, the coefficient of $\frac{\alpha^2}{\theta-1} \gg \alpha$, and thus $f(x \to \infty|\theta = 1 + \epsilon, \epsilon \to 0^+) \approx \frac{\alpha^2}{\theta-1} \frac{1}{x^{2\theta-1}} = \frac{\alpha^2}{\theta-1} \frac{1}{x^{1+2\epsilon}}$, indicating the scaling exponent $1 + 2\epsilon$ where $\epsilon \to 0^+$. In contrast, when $\theta \gg 1$, $f(x \to \infty|\theta \gg 1) \approx \frac{\alpha}{x^{\theta}}$, indicating the scaling exponent θ where $\theta \gg 1$. Thus, mathematically, we conclude the empirical power-law observations can come from the asymptotic behaviors of the stretched-exponential distribution. Besides, an interesting phenomenon is that as θ increases from 1, the distribution curve first becomes fatter and then back to previous states and further steeper and steeper. The above asymptotic analysis can be validated by the collapse of distributions with different θ values as shown in Fig. 4a.

5.1.3 Scalability. We compare the scalability of our model and the power-law (PL) method numerically. We generate N samples from power-law distribution configuration $\lambda(x|\beta = 0, \alpha = 1, \Delta = 50, \theta = 1)$, and by varying the N, we plot the average time consumed by two methods in Fig. 4b. The power-law method scales with complexity $\approx O(N^2)$. The rule-of-thumb upper-limit of the sample size for the power-law method is 10^5. The worse scalability of the PL method is due to the grid search of the Δ [5]. However, our method can be applied to much larger datasets with much faster speed. For instance, we get $\approx 2 * 10^3$ times faster when $N = 10^5$.

5.2 Real-world data analysis

5.2.1 Datasets. We validate our method by 16 real-world datasets from a variety of different human endeavors. According to the temporal nature of datasets, we classify them as cross-sectional data and dynamic data. The first eight datasets are from cross-sectional obervations:

(a) The number of occurrence of words in the novel Moby Dick by Herman Melville [16].

(b) The number of deaths by terrorist attacks worldwide from February 1968 to June 2006 [6].

(c) The number of mammals on Earth per taxonomic groups [20].

(d) The numbers of customers influenced by electrical blackouts in the US between 1984 and 2002 [16].

(e) The populations of US cities in the 2000 US Census [5].

(f) The acre sizes of wildfires occurring in US between 1986 and 1996 [16].

(g) The intensities of earthquakes occurring in California between 1910 and 1992 [16].

(h) The degree of actors in the movie-actor bipartite network [2].

The last eight ones are from dynamic records on human or social dynamics:

(i) The inter-event-time of adding consecutive friends in WeChat by an active user [25, 26].

(j) The IET of short-messaging from a mobile phone user [23].

(k) , *(l)* The response time of letter correspondence for Einstein and Freud during their whole life [18].

(m) The time intervals between an individual sending two consecutive e-mails during a 3 month period in a university [1, 21].

(n) The time interval of two retweets in a information cascade in Tencent Weibo [28, 29].

(o) The time interval of chatting behavior in an online group from Tencent QQ [32].

(p) The time interval of consecutive revision of one Wikipedia item [30].

We compile and public all the datasets (refer to Sec .7) for reproducibility, among which the last eight datasets serve as the first data collection on human and social dynamics.

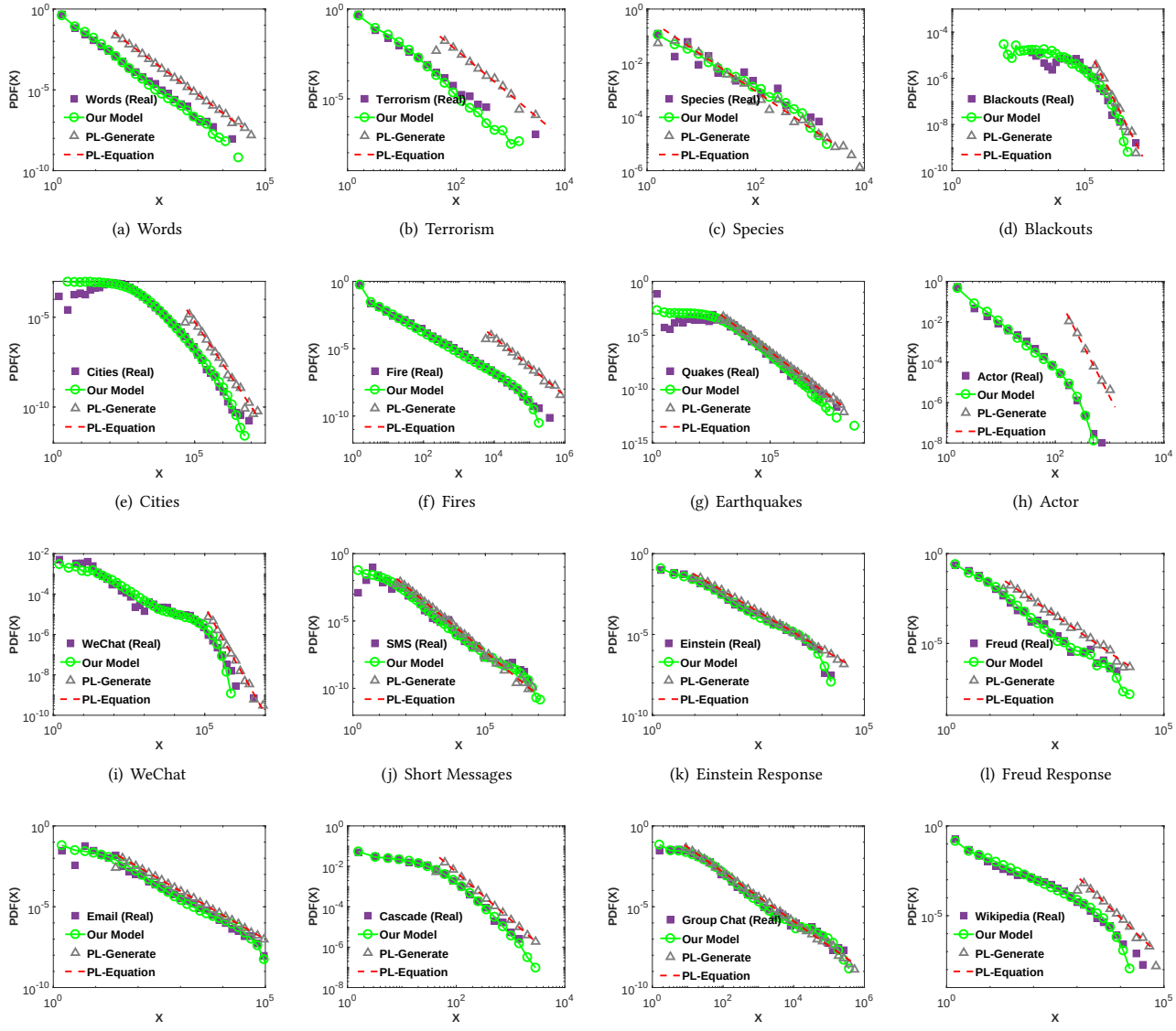


Figure 5: The $PDF(X)$ and their fitting results by our method and baselines [5] for 16 real-world datasets from a wide range of different disciplines. The real distributions are complex, and our method fits reality well. The first eight datasets are cross-sectional datasets: (a) The number of occurrence of words in the novel Moby Dick by Herman Melville. (b) The number of deaths by terrorist attacks worldwide from February 1968 to June 2006. (c) The number of mammals on Earth per taxonomic groups. (d) The numbers of customers influenced by electrical blackouts in the US between 1984 and 2002. (e) The populations of US cities in the 2000 US Census. (f) The acre sizes of wildfires occurring in US between 1986 and 1996. (g) The intensities of earthquakes occurring in California between 1910 and 1992. (h) The degree of actors in the movie-actor bipartite network. The last eight datasets are dynamic datasets: (i) The inter-event-time of adding consecutive friends in WeChat by an active user. (j) The IET of short-messaging from a mobile phone user. (k) and (l) The response time of letter correspondence for Einstein and Freud during their whole life. (m) The time intervals between an individual sending two consecutive e-mails during a 3 month period in a university. (n) The time interval of two retweets in a information cascade in Tencent Weibo. (o) The time interval of chatting behavior in an on-line group from Tencent QQ. (p) The time interval of consecutive revision of one Wikipedia item. Our model (green circles) fits all these datasets (purple squares) quite well, while the state-of-art method (dashed lines for the equation curve, triangles for the generated samples from the equation) shows large bias.

Table 2: Results of real-world datasets. Basic statistics of the 16 datasets, results of the PL method and the results of our method. With respect to KS-Dist error, our model captures the real datasets with much smaller error than the PL Method. The pdf described in PL method is $f(x) = \frac{\alpha_{PL}-1}{x_{min}}(\frac{x}{x_{min}})^{\alpha_{PL}}$.

	Real-World Data Statistics					PL Method			Our Method				
Dataset	N	$Min(X)$	$Max(X)$	$E[X]$	$Std(X)$	$\hat{x}_{min}$	$\hat{\alpha}_{PL}-1$	**KS-Dist**	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\Delta}$	$\hat{\theta}$	**KS-Dist**
(a) **Words**	18855	1	14086	11.14	148.33	26.00	0.93	0.960	$3.63e-04$	5.00	6.31	1.34	0.319
(b) **Terrorism**	9101	1	2749	4.35	31.58	50.00	1.52	0.992	$1.18e-10$	5.00	7.20	1.21	0.348
(c) **Species**	29	1	1425	148.41	324.35	2.00	0.36	0.208	$1.00e-03$	0.33	3.93	0.95	0.138
(d) **Blackouts**	211	1000	7500000	253868.68	610308.58	230000.00	1.27	0.725	$1.58e-06$	0.03	10000.00	0.80	0.108
(e) **Cities**	19447	1	8008654	9002.05	77825.05	52457.00	1.37	0.970	$7.99e-07$	0.40	723.77	0.92	0.033
(f) **Fire**	203785	0	412050	89.56	2098.73	6324.00	1.16	0.997	$3.53e-05$	0.53	0.14	1.00	0.249
(g) **Quakes**	19302	1.00	63095734.45	24537.21	563830.70	794.33	0.64	0.439	$1.96e-15$	0.37	521.91	0.92	0.095
(h) **Actor**	383640	1	646	3.83	10.42	162.00	4.21	0.999	$2.05e-02$	1000.00	14.54	2.83	0.388
(i) **WeChat**	973	0	4073278	57644.40	159193.93	122841.00	1.66	0.887	$1.22e-05$	0.11	30.00	1.12	0.076
(j) **SMS**	1692	0	4932276	16502.89	201848.27	45.00	0.62	0.556	$4.76e-07$	1.75	26.17	1.17	0.134
(k) **Einstein**	5943	0	18496	197.32	819.46	9.00	0.53	0.483	$5.09e-04$	10.00	18.55	1.62	0.076
(l) **Freud**	1190	0	7760	44.38	369.65	22.00	0.66	0.911	$3.06e-04$	10.00	12.83	1.51	0.157
(m) **Email**	9856	1	228965	711.70	5086.52	34.00	0.49	0.661	$6.57e-05$	7.37	38.05	1.43	0.121
(n) **Cascade**	3087	0	1586	52.42	102.59	49.00	1.36	0.720	$6.82e-04$	10.00	96.22	1.26	0.021
(o) **Group Chat**	1055	0	266831	2200.11	16078.36	8.00	0.52	0.245	$1.57e-05$	10.00	40.81	1.47	0.082
(p) **Wikipedia**	4660	1	29594.	311.58	1143.88	1153.00	1.32	0.940	$4.77e-04$	0.22	1.52	0.92	0.111

5.2.2 Results. We validate our method by answering if our model can capture all the empirical datasets. We compare our method to the state-of-the-art method developed in [5], denoted as PL model, which is widely used to fit the fat-tailed distributions which possibly follow power law.

Figure 5 plot the real datasets, fitting results by the PL model and our model. We find distributions of the real-world datasets are much more complex than a pure power-law distribution. For different data, the distribution of which exhibits different multi-scale complexity. However, the overlaps of our model (green circles) and the real data (purple squares) in all the plots indicate the good performance of model, even by visual inspection. According to the synthetic data analysis, the multi-scale complexities in distributions lead to serious overestimation of the x_{min} of the PL model, and we also observe these biases in the real datasets as shown in Fig. 5. And thus, the $f(x)$ learned by the PL model is systematic biased, as indicated by the discrepancies of the gray triangles representing the PL results and the purple squares representing real data.

We then conduct quantitative analysis. Given the real data $X = \{x_1, ..., x_n\}$, we learn the parameters Θ of the pdf $f(x|\Theta)$ modeled by the PL method and our method. We then generate the simulated data samples from our method. denoted as $X_{our} = \{x_1, ..., x_{n'}\}$, and $X_{PL} = \{x_1, ..., x_{n'}\}$ from PL method. We evaluate the fitting accuracy by the two-sample Kolmogorov-Smirnov distance (KS-Dist), i.e., KS-Dist = $max_x|\hat{F}_i(x) - F(x)|$, and the lower the better. The $F(x)$ is the non-parametric cumulative distribution learned from the real data, while the $\hat{F}(x)$ is the non-parametric cumulative distribution learned from the simulated datasets by method i. The two-sample Kolmogorov-Smirnov distance is widely used for this standard statistical test mission. In order to remove the errors due the small number of the generated samples, we set $n' = 10 * n$. We summarize the data and results in Table. 2. We find for all 16 datasets, our method get much lower error, namely KS-Dist, than the PL method, indicating the superiority of our method.

6 DISCUSSIONS

The design principle of our show-cased model is: keep simple, capture complex. Only one parameter Δ is used to capture the complexity in short-scale regime, one parameter β to capture the complexity in long-scale regime, and one parameter θ to encompass both power-law distribution and stretched-exponential distribution in the middle-scale regime. However, more efforts can be made by following our framework. For example, mixture heavy-tailed model, log-normal distribution can be further utilized. And we can expect much smaller errors between empirical data and the fitting results. However, no matter how complex the model is, modeling parameters should be interpretable. Further, prior knowledge on the parameters can be captured by a Bayesian framework. More real-world datasets should be examined. Some previous conclusions based on applying PL method on complex distributions should be re-examined.

7 CONCLUSIONS

In this paper, we find distributions in various empirical data, ranging from art, biology, physics, geology, social science to computer science, from cross-sectional observations to dynamic records, take on multi-scale complexities. We develop a dynamic framework to fit the complex distributions in the real world. By modeling the generative dynamics of the data, we extremely simplify the mathematical form of the model, but at the same time generate a wide range of complex distributions. Efficient inference method and data generative algorithm are provided. Rather than a black-box model, we explain the generative mechanisms of these complex distributions by a unified differential equation. We analyze the properties of our model by various synthetic datasets, and validate our model by a wide range of real-world datasets. Our model captures the complexity in all these data quite well. Our model potentially provides a framework to fit complex distributions in empirical data, and to understand their generative mechanisms. In short, we summarize our contributions as follows:

- *Unification power:* We propose a general model to fit various complex distributions in empirical data, together with inference and simulation algorithms.
- *Parsimony:* With four parameters, our model has a simple form to capture multi-scale complexities in empirical distributions.
- *Interpretability:* Our model is interpreted by a unified generative dynamic equation. All the parameters have clear physical meanings in the random network scenarios.
- *Usefulness:* Our model fits complex empirical dataset distributions in various disciplines accurately, and can be generalized to more complex cases in a principled way.

We open-source our code and datasets at www.calvinzang.com.

Acknowledgement

The authors thank anonymous reviewers for many useful discussions and insightful suggestions. This work was supported in part by National Program on Key Basic Research Project No. 2015CB352300, National Natural Science Foundation of China Major Project No. U1611461; National Natural Science Foundation of China No. 61772304, 61521002, 61531006, 61702296. Thanks for the research fund of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology, and the Young Elite Scientist Sponsorship Program by CAST.

REFERENCES

[1] Albert-Laszlo Barabasi. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435, 7039 (2005), 207–211.
[2] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.
[3] Anna D Broido and Aaron Clauset. 2018. Scale-free networks are rare. *arXiv preprint arXiv:1801.03400* (2018).
[4] Richard H Byrd, Jean Charles Gilbert, and Jorge Nocedal. 2000. A trust region method based on interior point techniques for nonlinear programming. *Mathematical Programming* 89, 1 (2000), 149–185.
[5] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review* 51, 4 (2009), 661–703.
[6] Aaron Clauset, Maxwell Young, and Kristian Skrede Gleditsch. 2007. On the frequency of severe terrorist events. *Journal of Conflict Resolution* 51, 1 (2007), 58–87.
[7] Luc Devroye. 1986. Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on Winter simulation*. ACM, 260–265.
[8] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. 1999. On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, Vol. 29. ACM, 251–262.
[9] Benjamin H Good, Michael J McDonald, Jeffrey E Barrick, Richard E Lenski, and Michael M Desai. 2017. The dynamics of molecular evolution over 60,000 generations. *Nature* 551, 7678 (2017), 45.
[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.
[11] Lei Guo, Enhua Tan, Songqing Chen, Zhen Xiao, and Xiaodong Zhang. 2008. The stretched exponential distribution of internet media access patterns. In *Proceedings of the twenty-seventh ACM symposium on Principles of distributed computing*. ACM, 283–294.
[12] R Dean Malmgren, Daniel B Stouffer, Adilson E Motter, and Luís AN Amaral. 2008. A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences* 105, 47 (2008), 18153–18158.
[13] Michael Mitzenmacher. 2004. A brief history of generative models for power law and lognormal distributions. *Internet mathematics* 1, 2 (2004), 226–251.
[14] Kevin P. Murphy. 2014. Machine learning, a probabilistic perspective. (2014).
[15] Mitchell G Newberry, Christopher A Ahern, Robin Clark, and Joshua B Plotkin. 2017. Detecting evolutionary forces in language change. *Nature* 551, 7679 (2017), 223.
[16] Mark EJ Newman. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary physics* 46, 5 (2005), 323–351.
[17] John Nolan. 2003. *Stable distributions: models for heavy-tailed data*. Birkhauser New York.
[18] Joao Gama Oliveira and Albert-László Barabási. 2005. Human dynamics: Darwin and Einstein correspondence patterns. *Nature* 437, 7063 (2005), 1251–1251.
[19] Douglas Reynolds. 2015. Gaussian mixture models. *Encyclopedia of biometrics* (2015), 827–832.
[20] Felisa A Smith, S Kathleen Lyons, SK Ernest, Kate E Jones, Dawn M Kaufman, Tamar Dayan, Pablo A Marquet, James H Brown, and John P Haskell. 2003. Body mass of late Quaternary mammals. *Ecology* 84, 12 (2003), 3403–3403.
[21] Alexei Vázquez, Joao Gama Oliveira, Zoltán Dezsö, Kwang-Il Goh, Imre Kondor, and Albert-László Barabási. 2006. Modeling bursts and heavy tails in human dynamics. *Physical Review E* 73, 3 (2006), 036127.
[22] G West. 2017. Scale: The universal laws of growth, innovation, sustainability and the pace of life in organisms and companies. (2017).
[23] Ye Wu, Changsong Zhou, Jinghua Xiao, Jürgen Kurths, and Hans Joachim Schellnhuber. 2010. Evidence for a bimodal distribution in human communication. *PNAS* 107, 44 (2010), 18803–18808.
[24] Manzil Zaheer, Chun-Liang Li, Barnabás Póczos, and Ruslan Salakhutdinov. 2017. GAN Connoisseur: Can GANs Learn Simple 1D Parametric Distributions? (2017).
[25] Chengxi Zang, Peng Cui, and Christos Faloutsos. 2016. Beyond Sigmoids: The NetTide Model for Social Network Growth, and Its Applications. In *Proceedings of the 22Nd ACM SIGKDD (KDD '16)*. ACM, 2015–2024.
[26] Chengxi Zang, Peng Cui, Christos Faloutsos, and Wenwu Zhu. 2017. Long Short Memory Process: Modeling Growth Dynamics of Microscopic Social Connectivity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 565–574.
[27] Chengxi Zang, Peng Cui, Christos Faloutsos, and Wenwu Zhu. 2018. On Power Law Growth of Social Networks. *IEEE Transactions on Knowledge and Data Engineering* (2018).
[28] Chengxi Zang, Peng Cui, Chaoming Song, Christos Faloutsos, and Wenwu Zhu. 2017. Quantifying Structural Patterns of Information Cascades. In *Proceedings of the 26th International Conference on WWW Companion*. 867–868.
[29] Chengxi Zang, Peng Cui, Chaoming Song, Christos Faloutsos, and Wenwu Zhu. 2017. Structural patterns of information cascades and their implications for dynamics and semantics. *arXiv preprint arXiv:1708.02377* (2017).
[30] Yilong Zha, Tao Zhou, and Changsong Zhou. 2016. Unfolding large-scale online collaborative human dynamics. *Proceedings of the National Academy of Sciences* 113, 51 (2016), 14627–14632.
[31] Tianyang Zhang, Peng Cui, Christos Faloutsos, Yunfei Lu, Hao Ye, Wenwu Zhu, and Shiqiang Yang. 2016. Come-and-go patterns of group evolution: A dynamic model. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1355–1364.
[32] Tianyang Zhang, Peng Cui, Chaoming Song, Wenwu Zhu, and Shiqiang Yang. 2016. A multiscale survival process for modeling human activity patterns. *PloS one* 11, 3 (2016), e0151473.