On Power Law Growth of Social Networks

Chengxi Zang, Peng Cui, *Member, IEEE,* Christos Faloutsos, and Wenwu Zhu, *Fellow, IEEE*

Abstract—What is the growth dynamics of social networks, like Facebook or WeChat? Does it truly exhibit exponential early-growth, as predicted by the celebrated models, like the Bass model? How about the dynamics of links, for which there are few published models? For the first time, we examine the growth of WeChat which is the largest online social network in China, together with several other real social networks. We observe Power-Law growth dynamics for both nodes and links, a fact that breaks the textbook models featuring Sigmoid curves. We propose NETTIDE, along with differential equations for the growth of nodes and links. Our model fits the growth dynamics of real social networks well; it encompasses many traditional growth dynamics as special cases; while remaining parsimonious in parameters. The NETTIDE for link growth is the first one of its kind, accurately fitting real data, and capturing densification phenomenon. We further formulate two stochastic generators, which interpret the growth of nodes and links through survival analysis and micro-level interactions within a social network respectively. The proposed generators reproduce realistic growth dynamics of social networks. When applied on the WeChat data, our NETTIDE forecasted $\geq$ 730 days ahead with 3% error.

Index Terms—Social Network, Social Dynamics, Growth Dynamics, Power-Law Growth, Link Growth, Stretched-Exponential Growth.

1 INTRODUCTION

GROWTH dynamics of social systems, ranging from social networks [43], [44], social groups [46] to information cascades [41], [45], occupies a central place in understanding social dynamic phenomena. However, the lack of data which documented the evolution of large social systems in real world prohibits us from understanding intrinsic mechanisms governing their growth dynamics. Thus, it's difficult to answer common questions like: How many members will Twitter have next month? How many friendship links will WeChat[1] (or Facebook, or Google-plus) have next year? The count of members of a network (or belief, or religion, or epidemic) is of vital importance (growth of social products, provisioning, social implications of policy changes, etc) and has been studied extensively (see section 2). The count of links has attracted less interest, although it is also important (well connected nodes in, say, FaceBook, are less inclined to churn; well connected neurons in a brain possibly indicate resistance to Alzheimer's disease, etc).

Researchers from multiple disciplines have studied network growth phenomenon for decades [7], [20], [21], [24], [26], [34], and have achieved significant advancement towards understanding the generation of scale-free networks, the densification of network links, the shrinking diameters and so on. Network growth models include the celebrated Barabási-Albert model and its variants [7], [8], [10] - they all assume uniform growth of nodes. The Bass model [29] and the Susceptible-Infected (SI) model [3], produce Sigmoid growth curve with exponential growth at early stage. None of them studies the growth of links over time.

In short, the focus of this paper is to answer these three questions on growth dynamics of social networks:

1) How does the number of nodes $n(t)$ grow over time?
2) How does the number of links $e(t)$ grow over time?
3) Can we generate realistic $n(t)$ and $e(t)$?

The reader may think that, at least the first question, already has an answer: Sigmoid growth (which is the solution to the SI and the Bass model). However, reality disagrees, exhibiting power-law growth, instead, as shown in Figures 1a-b. Specifically, we examine the evolving processes of four real social networks, including WeChat, arXiv [1], Enron [19] and Weibo [42], respectively representing online social networks, co-authorship social networks, enterprise social networks and information cascading networks. Taking WeChat for instance, we study its detailed evolution from zero to 300 million nodes and 4.75 billion links, spanning two years. We find that although the growth curves of the four social networks have different shapes, they all follow a power-law like growth pattern. Specifically, we find the growth dynamics of WeChat follows power-law growth with exponent 2.15 for nodes and 3.01 for links (Fig 1a), and the growth dynamics of arXiv follows power-law like growth before hitting the plateaus (Fig 1b). These observations go beyond our traditional expectations of exponential or uniform growth dynamics of social networks, or diffusion of innovations phenomena.

Since Sigmoid and related growth models are contradicted by reality, we need a more realistic one. We shoot for a model to fulfill the following characteristics which good models should have:

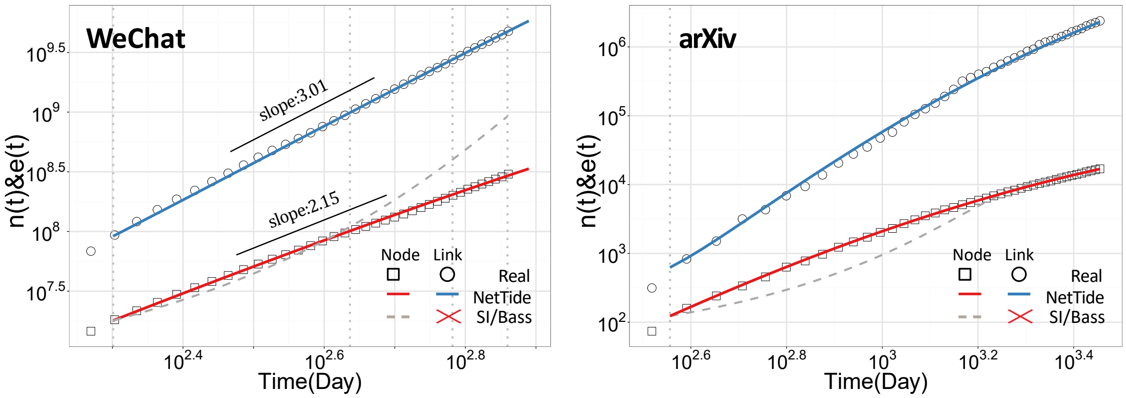
1) *Parsimony:* The model should have as few parameters as necessary, and still generate power-law growth.
2) *Generality:* The model should be general, better to encompass traditional growth models like Bass, SI and Log-Logistic growth, etc., as special cases.
3) *Link growth:* The model should be able to capture the growth dynamics of links.
4) *Intuition:* The model should easy to interpret.
5) *Generator:* The population dynamic model should correspond to microscopic generators which produce realistic growth dynamics at microscopic level.

We propose a novel dynamic model, named NETTIDE, for network growth. The NETTIDE model consists of two components, NETTIDE-Node and NETTIDE-Link for nodes and links respectively. As we show later, our NETTIDE achieves the afore-

• C. Zang, P. Cui, and W. Zhu are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. E-mail: zangcx13@mails.tsinghua.edu.cn, {cuip, wwzhu}@tsinghua.edu.cn
• C. Faloutsos is with the Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15217. E-mail: christos@cs.cmu.edu

Manuscript received April 19, 2005; revised August 26, 2015.

1. www.wechat.com/en/



(a) WeChat (b) arXiv

Fig. 1. *Reality disobeys Sigmoid curves*: We find power-law growth dynamics for WeChat (a), and arXiv (b) - nodes over time (squares □), well fitted by our proposed NETTIDE-Node (solid red line), but not by the SI or Bass model (dashed gray line). The link count over time (circles ○) and our fitting model (NETTIDE-Link - in solid blue). Notice that there is no competitor for link dynamics (SI or Bass is crossed out). All axes are in log scale.

mentioned characteristics, and it matches the behavior of several, disparate real social networks.

Our model can be applied for prediction, clustering, and outlier detection [18] on growth dynamics. For example, we show that NETTIDE is able to forecast the growth of WeChat almost 2 years in the future (730 days), with $\approx 3\%$ error, which is impressive given the fact that social systems are complex and non-linear. As points of reference, most forecasting works usually do just one step look-ahead [27], [32], rather than trying to capture and then predict the long-term dynamics of complex systems.

The intuition of NETTIDE let us formulate two stochastic generators, i.e., NETTIDE-Survival and NETTIDE-Process, both of which generate realistic stochastic dynamics of nodes and links. The NETTIDE-Survival connects the NETTIDE with the survival analysis framework by modeling the hazard rates of nodes and links. The NETTIDE-Process interprets the NETTIDE as a collective dynamics derived from micro-level stochastic interactions within a network. Through extensive numerical simulations, both stochastic versions generate realistic growth dynamics.

In summary, the contributions of this work and the advantages of NETTIDE are as follows:

- *Novel model* NETTIDE: It is parsimonious, it generalizes many growth models as special cases, it provides the first-ever differential equation for link growth, and it is intuitive.
- *Accuracy:* NETTIDE fits the growth of several, diverse, real networks accurately.
- *Usefulness:* NETTIDE gives excellent forecasting, down to *3%* error, for almost *2 years* ahead in the future.
- *Generators:* NETTIDE has two *stochastic* generators, i.e., NETTIDE-Survival and NETTIDE-Process, which generate realistic stochastic growth dynamics of social networks.

Several of the datasets are public [1], [19]; our code is open-sourced at github.com/calvin-zcx/NetTide

The outline of the paper is the typical one: we give the survey (Section 2), the proposed method (Section 3), experiments (Section 4), and conclusions (Section 5).

2 RELATED WORK

As the investigated problem is closely related to network evolution, growth models, and human dynamics, we mainly review the related works in these three fields.

2.1 Evolving network

The pioneering studies in evolving network have revealed that the growth process of a real network plays a vital role in shaping its structure, like the power-law distribution of degree [7], shrinking diameters [26], densification growth [26] and so on. However, all these evolving network models assume that the dynamics of the node growth process are uniform, like the Barabási-Albert model (BA) model [7] and its variants [10]. Some other works empirically exploit node growth dynamics as input, and do not aim to find the patterns of node growth dynamics [9], [15], [17], [22], [24], [26].

In literature, the growth dynamics of links are largely ignored. A few works show the double preferential attachment or the random attachment [2], [8], [14] of the internal links may make the network more homogeneous, where the growth rate of links are also assumed to be uniform. Recently, the effects of information diffusion on link creation are studied in [4], [39], and [12] proposed the multivariate Hawkes process model (Coevolve) to capture the microscopic evolution of the linking and information diffusion. However, it suffers from two issues: computation time is prohibitive, being $O(N^2)$ where N represents the number of events as mentioned in [12] ; being based on a Hawkes process, it can NOT generate power-law growth with the observed exponents $(2.15, 3.01$ in Fig 1). See discussion on Hawkes process, in section 2.2. Temporal and spatial correlations of the growth of social networks are studied in [37], [48].

2.2 Growth models

The growth models [5] are discussed in a wide range of fields. The most classical models on growth phenomenon are Susceptible Infected (SI) model in epidemiology [3] and the Bass model [29] on the diffusion of innovations. They generate S-shaped curve with exponential growth at early stage for the growth of infected nodes. They also provide intuitive explanation for the microscopic infection process in a mean-field form. The exponential growth induced by the constant infection rate is against the intuition of the forgetting nature of human [6], [35], or the fizzling patterns in the social networks [28]. Models like PhoenixR [13] tried to introduce fizzling mechanism based on Susceptibel Infected Recovered (SIR) model [3]. The variances of SIR models are further

TABLE 1
Capabilities of models. Only our model meets all specs.

	Net Growth			Growth phenomenon					Our model
Capability	BA	FF	Coevolve	SI	BASS	CS	SpikeM	PhoenixR	**NetTide**
Exponential growth			✓	✓	✓	✓	✓	✓	✓
Power-law growth with arbitrary exponent									✓
Differential equation for $n(t)$				✓	✓	✓	✓	✓	✓
Closed form $n(t)$				✓	✓				✓
Differential equation for $e(t)$			✓						✓
Microscopic Generators	✓	✓		✓					✓

developed in [36], [46], etc. However, the constant recovery rate based on SIR can not slow the exponential growth down to the power law growth. In all, all the above models cannot generate power-law growth as we observed in the real social network data.

Recently, studies based on a kind of self-excited point process, i.e. Hawkes processes (HP) [16], are introduced to capture the growth and diffusion phenomena, which can be viewed as the endogenous branching process (BP) with the exogenous immigration process [23], like Crane-Sornette (CS) model [11], SpikeM [33]. The HP and its variants above can generate growth patterns in three regimes: exponential growth in super-critical regime like SpikeM, the power law growth with exponent < 1 for rate and exponent < 2 for the cumulative count like CS model, and the growth dying out quickly in sub-critical regime. Thus, all the above models cannot generate power-law growth with arbitrary exponent, like 2.15 we observed in WeChat social network. In addition, none of previous models describe the growth dynamics of links.

2.3 Human dynamics

The dynamics of human beings usually exhibit bursty behaviors and fat-tailed interevent time (IET) distribution [6]. Power-law distribution of IET are explained by priority queue model [38] and modulated Poission model [30]. Recently, multiscale nature of IET distribution [44], [47] have been found, including bimodal distribution [40] at long-scale, and quick actions at short-scale [47]. In all, fizzling temporal correlations of human behaviors are ubiquitous, serving as one of the fundamental pillars of modeling human behavior.

We summarize the relative advantages and failures of all above models in Table 1. Only our NETTIDE model meets all the advantages.

3 THE NETTIDE MODEL

3.1 Preliminaries

Traditional models, like the SI model or the Bass model, fail to capture the power-law growth as shown in Figure 1. The SI model is powerful, intuitive, and heavily used in numerous fields, either as-is or with tiny modifications (like the Bass model, that adds a noise term). However, it fails qualitatively to match the real-world data of Figure 1a and 1b: it can only generate Sigmoid growth over time, which lead to exponential early-growth - not power-law growth.

Some reasonable attempts. Maybe we should vary the infectivity factor β, say, decaying over time (possibly because the novelty wears off). An obvious way to show diminishing interest would be exponential decay, i.e., $\beta = \beta_0 * \exp(-\xi t)$ where ξ is the half-life of the radioactive-like decay of enthusiasm:

TABLE 2
Symbols and Definitions

Symbols	Definitions
N	Number of the total population
$n(t)$	Cumulative number of users by time t
$dn(t)/dt$	Number of new users at time t
$e(t)$	Cumulative number of links by time t
$de(t)/dt$	Number of new links at time t
β	Maximum growth rate of nodes
θ	Temporal fizzling exponent
β'	Maximum linking rate
γ	The scaling sparsity exponent
α	The linear sparsity coefficient

- Attempt 1: *Radioactive decay:*

$$\frac{dn(t)}{dt} = \beta_0 \exp(-\xi t) n(t) \, (N - n(t))$$

But above combination cann't give the the power-law growth of Figure 1.

The growth rate should depend on $n(t)$ (the infected ones) as well as on the susceptibles ($N - n(t)$) - but maybe not linearly? Maybe not all susceptibles are available (e.g., some of them are taking precautions against the infection) and/or not all infected ones are actually active (e.g., some of them stay home). With less-than-full participations, the equation becomes:

- Attempt 2: *Partial participation(s)*:
  We tried $n(t)^\zeta$, as well as $(N - n(t))^\psi$, where $\zeta < 1, \psi < 1$ try to model the less-than-full participation:

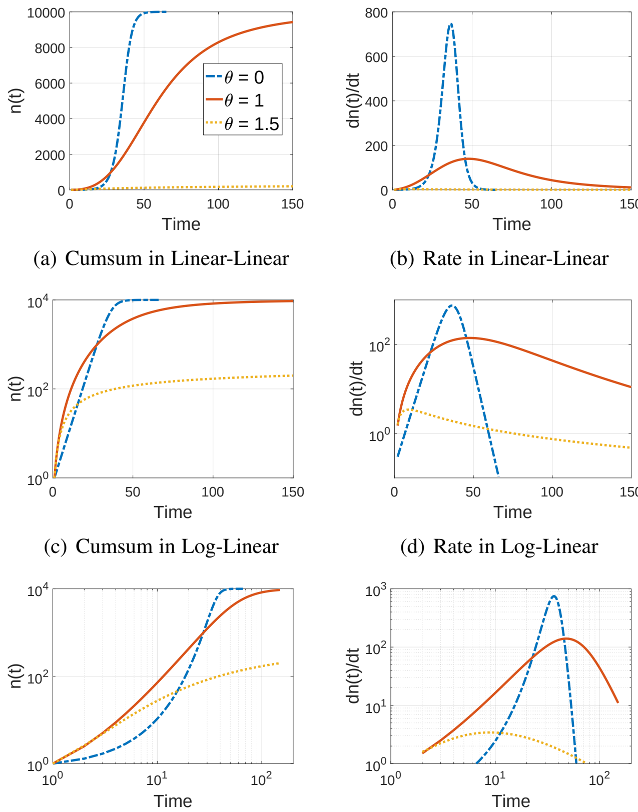
$$\frac{dn(t)}{dt} = \beta n(t)^\zeta \, (N - n(t))^\psi$$

This model has been widely used to capture the rate of a chemical reaction, but the $n(t)^\zeta$ and $(N - n(t))^\psi$ participatants are hard to interpret to microscopic process. We will show that even without any additional parameter our model fits power-law growth reality quite well, and with additional one parameter it generates a wide range of growth dynamics.

3.2 NETTIDE-Node

It turns out that full participations but with decay of the infectivity/enthusiasm β capture the reality. A good model should be intuitive - why would human interest decay?Decays, especially scaling or power-law decays, have been observed in social interactions (email response times etc, as we mentioned in the related work section); as well as in the theory of random walks (the time

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2018.2801844, IEEE Transactions on Knowledge and Data Engineering



(a) Cumsum in Linear-Linear (b) Rate in Linear-Linear

(c) Cumsum in Log-Linear (d) Rate in Log-Linear

(e) Cumsum in Log-Log (f) Rate in Log-Log

Fig. 2. NETTIDE-*Node generates various growth dynamics*. The cumulative growth curves with different parameters on different scales are shown in a, c, e, and corresponding rate curves are shown in b, d, f. *Sigmoid growth* with *exponential early growth* ($\theta = 0, N = 10^4, \beta = 3 \times 10^{-5}$), *Log-Logistic growth* with *Power-Law early growth* ($\theta = 1, N = 10^4, \beta = 3 \times 10^{-4}$), and *Stretched-Logistic growth* with *Stretched-Exponential early growth* ($\theta = 1.5, N = 10^4, \beta = 4.5 \times 10^{-4}$) are in blue dashed, red solid and yellow dotted line respectively.

between zero-crossings follows power law with exponent -1.5). The power law decays are justifiable and intuitive to capture human decision-making process and collective relaxation dynamics of social system.

It is the decay (or fizzling) mechanism leading to the power-law growth, as shown in Fig. 1 and even more general case stretched-exponential growth as shown in Fig. 2. Next, we give the details and the proofs for the growth of nodes (NETTIDE-Node model). Further in the next subsection, we work on the links (NETTIDE-Link model). We summarize the symbols in table 2.

As we said, our NETTIDE-Node is governed by the differential equation below:

$$\frac{dn(t)}{dt} = \frac{\beta}{t^{\theta}} n(t)(N - n(t)) \tag{1}$$

A social network with a large population $n(t)$ has a propensity to attract more nodes in the early stage. As the population who can join the social network is limited, its growth will be constrained by the decreasing number of potential nodes $(N - n(t))$, especially at the saturation stage. This is a natural phenomenon and has been observed in numerous disciplines, from the law of mass action in chemistry to model the rate of a chemical reaction, to the spreading of disease between the susceptible and the infected in epidemics. The term $\frac{\beta}{t^{\theta}}$ $(t > 0)$ is the fizzling infection/excitement rate since the inception of the social network. That is, people have decaying excitement to infect their friends to join a social network. It is exactly the scaling exponent θ of the decay that leads to various growth dynamics, including the power law growth as special case (when there is no θ in Eq. 1, or equivalently $\theta = 1$.). We refer to θ as temporal fizzling exponent.

Next, we give the proofs that (a) our NETTIDE-Node model can indeed lead to power law growth, (b) it includes the sigmoid models (SI etc.) as special cases, and (c) it is capable of generating a wide range of growth dynamics.

Lemma 1. *When $\theta = 1$,* NETTIDE-*Node follows* **Log-Logistic growth** *dynamics, shown in equation (3), which approximates the* **Power-Law growth** *shown in equation (6) with exponent βN when $n(t) \ll N$.*

Proof. When $\theta = 1$, the NETTIDE-Node leads to

$$\frac{dn(t)}{dt} = \frac{\beta}{t} n(t)(N - n(t)) \tag{2}$$

As this is a separable differential equations, we can separate $n(t)$ term and t term to do the integral separately, and then get

$$n(t) = N \frac{\lambda_0 \exp\{\int_{t_0}^{t} \frac{\beta N}{\mu} d\mu\}}{1 + \lambda_0 \exp\{\int_{t_0}^{t} \frac{\beta N}{\mu} d\mu\}} = N \frac{\lambda_0 (\frac{t}{t_0})^{\beta N}}{1 + \lambda_0 (\frac{t}{t_0})^{\beta N}} \tag{3}$$

where

$$\lambda_0 = \frac{n_0}{N - n_0} \tag{4}$$

and n_0 is the total number of nodes in the initial time t_0 of the system. If $n(t) \ll N$,

$$\frac{dn(t)}{dt} \approx \frac{\beta N}{t} n(t) \tag{5}$$

leads to

$$n(t) = n_0 (\frac{t}{t_0})^{\beta N} \tag{6}$$

which shows power law growth with exponent βN. □

The illustrations of the Log-Logistic growth, in red solid curve, are shown in Fig. 2. A stretched S-shaped cumulative growth curve is shown in Fig. 2a, compared with the sigmoid curve in blue dashed line. The power-law growth (red solid curve) exhibits sub-linear growth on log-linear scale shown in Fig. 2c, and features linear growth on log-log scale shown in Fig. 2e, indicating power-law rise and fall patterns for rate, before and after the inflection point respectively, compared with Sigmoid curve in blue (Figs. 2bd&f).

Lemma 2. *When $\theta \neq 1$,* NETTIDE-*Node follows growth pattern as in equation (7). When $n(t) \ll N$, the growth at early times behaves as equation (8).*

We name the newfound growth dynamics in equation (7) as **Stretched-Logistic growth**, and equation (8) as **Stretched-Exponential growth** at early stage.

Proof. When $\theta \neq 1$, the deviation procedures of equation (7) and the initial growth (8) are similar with Proof in Lemma 1. We get:

$$\begin{aligned} n(t) &= N \frac{\lambda_0 \exp\{\int_{t_0}^{t} \frac{\beta N}{\mu^{\theta}} d\mu\}}{1 + \lambda_0 \exp\{\int_{t_0}^{t} \frac{\beta N}{\mu^{\theta}} d\mu\}} \\ &= N \frac{\lambda_0 \exp\{\frac{\beta N}{1-\theta}(t^{1-\theta} - t_0^{1-\theta})\}}{1 + \lambda_0 \exp\{\frac{\beta N}{1-\theta}(t^{1-\theta} - t_0^{1-\theta})\}} \end{aligned} \tag{7}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2018.2801844, IEEE Transactions on Knowledge and Data Engineering

where λ_0 is defined in Lemma 1. When $n(t) \ll N$, the initial growth behaves as:

$$\begin{aligned} n(t) &= n_0 \exp\{\int_{t_0}^{t} \frac{\beta N}{\mu^\theta} d\mu\} \\ &= n_0 \exp\{\frac{\beta N}{1-\theta}(t^{1-\theta} - t_0^{1-\theta})\} \end{aligned} \tag{8}$$

Now, we show the reason why we name it Stretched-Logistic. It is worth recalling that if one random variable (r.v.) follows the Log-Logistic distribution, then its logarithm follows a logistic distribution. Following the naming rule of the Log-Logistic distribution, we then will show that if a r.v. T follows the Stretched-Logistic distribution as

$$P_T\{T \leq t\} = \frac{1}{Z_T} \frac{\lambda \exp\{\frac{\beta N}{1-\theta}(t^{1-\theta} - t_0^{1-\theta})\}}{1 + \lambda \exp\{\frac{\beta N}{1-\theta}(t^{1-\theta} - t_0^{1-\theta})\}} \tag{9}$$

, where Z_T is the normalization factor and λ is a constant, then its integral of fizzling effect $X = \int_{t_0}^{T} t^{-\theta} dt = \frac{T^{1-\theta}}{1-\theta} - \frac{t_0^{1-\theta}}{1-\theta}$ shall follow the Logistic distribution. When $\theta < 1$, for any $x \geq \frac{t_0^{1-\theta}}{\theta-1}$,

$$\begin{aligned} P_X\{X \leq x\} &= P_T\{\int_{t_0}^{T} t^{-\theta} dt \leq x\} \\ &= P_T\{\frac{T^{1-\theta}}{1-\theta} - \frac{T_0^{1-\theta}}{1-\theta} \leq x\} \\ &= P_T\{T \leq [(x + \frac{t_0^{1-\theta}}{1-\theta})(1-\theta)]^{\frac{1}{1-\theta}}\} \\ &= \frac{1}{Z_T} \frac{\lambda \exp\{\beta N x\}}{1 + \lambda \exp\{\beta N x\}} \end{aligned}$$

which shows that X follows Logistic distribution. When $\theta > 1$, for any $x < \frac{t_0^{1-\theta}}{\theta-1}$, similar procedures as above can prove that X follows Logistic distribution. □

Lemma 3. *When $\theta = 0$, the* NETTIDE*-Node follows the* **Sigmoid (or Logistic) growth** *dynamics, e.g. SI model, which is a special case of Lemma (2). When $n(t) \ll N$, the Logistic growth approximates to the* **Exponential growth** *as:*

$$n(t) = n_0 \exp\{\beta N (t - t_0)\} \tag{10}$$

Proof. Replace the θ in Equation (7) and Equation (8) with 0. Usually, the sigmoid function refers to a simplified case of logistic function. Here, we use sigmoid and logistic interchangeably to indicate S-shaped growth curve with exponential early growth. □

Justification of the NETTIDE-Node:

- *Temporal fizzling*. Instead of capturing the temporal fizzling effect of each individual by $\frac{\beta}{(t-t_i)^\theta}$, where t_i is the time of i entering the system, we describe the fizzling growth of the system by $\frac{\beta}{t^\theta}$, where t is the time tick since the inception of the whole system. Because the models capture the integral of individual decay like $\frac{dn(t)}{dt} = n(t_0) + \sum_{t_i \leq t} \mu_i \frac{1}{(t-t_i)^\theta}$ can only generate exponential growth or power law growth with exponent < 2 (as discussed in related work section). It fails the reality (non-exponential growth, or power law growth with arbitrary exponent like ≥ 2). In contrast, our NETTIDE-Node fits the real data very well (in Experiment section), and can encompass a large range of growth patterns: power law early-growth with arbitrary exponent, the general form stretched-exponential early-growth, and the exponential early-growth as a special case (illustrated in Fig. 2).

3.3 NETTIDE-Link

The growth of social network can never be limited to nodes only. No such differential equations to describe the growth dynamics of links do exist before. Here, we introduce NETTIDE-Link to characterize link growth dynamics. We assume that there exists underlying organizational structure as the context of social network formation and growth. For example, the formation and growth of co-author social networks is constrained by the organizational structures such as mentor-students and researcher-collaborators structures. Hence, we need to take into account the characteristics of the underlying organizational structure when modeling the network growth. We define the underlying organizational structure as graph G_0, and the linking process is described as follow: for each existing node i, i tries to link to his already existing neighbor j in G_0 . If there is a link already being there, then nothing happens. If the link from i to j has not been established yet, i tries to link j with rate β' over the temporal fizzling term t^θ. The arrival of new nodes will bring a constant number of external links. The NETTIDE-Link summarizes above linking process:

$$\frac{de(t)}{dt} = \frac{\beta'}{t^\theta} n(t)(\alpha(n(t)-1)^\gamma - \frac{e(t)}{n(t)}) + 2\frac{dn(t)}{dt} \tag{11}$$

Justification of the NETTIDE-Link:

- *External links.* $2\frac{dn(t)}{dt}$ captures the process where a newly-arriving node bring two new links because we treat a link as bidirectional link. The assumption is that we treat the first link of each newly-arriving node as the external link. Also we can elaborate on it, like treating the first m links of the newly-arriving node being made at the same time.
- *Internal links.* Internal links are built between the already-existing nodes, and thus give rise to the densification phenomenon. For each existing node, he/she tries to link the existing neighbors in G_0 which have not being linked. Because of the organizational structure, less-than-full existing nodes can be accessed and $\alpha(n(t)-1)^\gamma$ captures the average accessible existing neighbors. The term $\frac{e(t)}{n(t)}$ is the average number of already linked neighbors to be excluded. The $\frac{\beta'}{t^\theta}$ captures the fizzling linking rate.
- *Densification.* By empirical analysis in experiment section, the link equation captures the densification power law by the power-law sparsity exponent γ. The densification power law between links and nodes are $1+\gamma$. For example, if we model the G_0 by the Kronecker graph model [25], we get $\gamma = \log E / \log N$.

3.4 Stochastic generators

Here we formulate two stochastic generators of NETTIDE, i.e., NETTIDE-Survival and NETTIDE-Process, which generate realistic stochastic growth dynamics of social networks.

NETTIDE-Survival. We first show NETTIDE can be explained in the survival analysis framework by defining hazard rates for node and link growth respectively.

- *Hazard rate of node growth.* The instantaneous rate that an individual adoption of, say, WeChat, will be made at time epoch t given that this individual has yet been a registered user is

$$\lambda_n(t) = \frac{\beta}{t^\theta} n(t). \tag{12}$$

Let $F(t)$ be the fraction of individuals who have been a registered user by time t and $f(t)$ denotes the derivative of $F(t)$. Then, $\lambda_n(t) = \frac{f(t)}{1-F(t)} = \frac{\beta}{t^\theta}n(t)$. Notice that $F(t) = \frac{n(t)}{N}$ and $f(t) = \frac{dn(t)}{N\times dt}$, thus we connect the NETTIDE-Node equation (1) and the hazard rate function (12) by

$$\begin{aligned}\frac{dn(t)}{dt} &= f(t)\times N \\ &= \lambda_n(t)\times(1-F(t))\times N \\ &= \frac{\beta}{t^\theta}n(t)(N-n(t)).\end{aligned} \tag{13}$$

Indeed, the hazard rate function in survival analysis framework and the (Log/Fizzle-) Logistic framework are connected by equation (13), based on the assumption that the potential population N is constant.

- *Hazard rate of link growth.* The instantaneous rate of the establishment of the external links are two times of the instantaneous rate of nodes' adoption rate. As for the internal links, the instantaneous rate that an internal link will be made at t given that this link has yet been built is

$$\lambda_e(t) = \frac{\beta'}{t^\theta}. \tag{14}$$

Following similar analysis procedure, let $F_e(t)$ be the fraction of internal links which have been built by time t and $f_e(t)$ is the derivative of $F_e(t)$. Then,

$$\begin{aligned}\lambda_e(t) &= \frac{f_e(t)}{1-F_e(t)} \\ &= \frac{de(t)}{(\alpha n(t)(n(t)-1)^\gamma - e(t))dt} \\ &= \frac{\beta'}{t^\theta},\end{aligned} \tag{15}$$

where $\alpha n(t)(n(t)-1)^\gamma - e(t)$ is the total number of potential links to be built by time t minus the links already built by t. Thus we connect the NetTide-Link equation (11) and the hazard rate function (14) by equation (15).

Given the input parameters $(\beta, \theta, N, \beta', \alpha, \gamma)$, the NETTIDE-Survival generate stochastic growth dynamics $n(t)$ and $e(t)$ from hazard rates as follows:

- *Node growth.* At each time epoch t, each of $N-n(t)$ survival nodes is selected randomly to join with probability $\lambda_n(t)h + o(h)$ during the time interval $(t, t+h)$ where $h \to 0$. If success, we increase $n(t)$ by one and $e(t)$ by two.
- *Link growth.* For the $\alpha n(t)(n(t)-1)^\gamma - e(t)$ survival links by t, each of them is selected with probability $\lambda_e(t)h + o(h)$ during the time interval $(t, t+h)$ where $h \to 0$. If success, we increase $e(t)$ by two.

Extensions of NETTIDE-Survival. We could extend our model to capture more realistic situations. For instances, hazard function of node can have poisson term λ_0 for $\lambda_n(t) = \lambda_0 + \frac{\beta}{t^\theta}n(t)$ to capture the intrinsic propensity for an individual to make an adoption. The time-lag effect at the burning period can be captured by $\lambda_n(t) = \frac{\beta}{(t+\Delta)^\theta}n(t) = \frac{\beta/\Delta^\theta}{(1+t/\Delta)^\theta}n(t)$. By introducing the sinusoidal function with specific configurations, we can incorporate the periodicity into NETTIDE. For clarity, we do not elaborate on these situations.

NETTIDE-Process. We further interpret the node and link growth dynamics from the perspective of micro-level stochastic interactions within a network. Under the mean-field assumption that each individual is apportioned uniformly to the susceptible neighbors, we can decompose the aggregate hazard rate into the micro-level pairwise infection rate:

$$\begin{aligned}\frac{dn(t)}{dt} &= \lambda_n(t)\times(N-n(t)) \\ &= \frac{\lambda_n(t)}{\langle k\rangle}\langle k\rangle N(1-F(t)) \\ &= \frac{\beta N}{\langle k\rangle t^\theta}n(t)\langle k\rangle(1-F(t)),\end{aligned} \tag{16}$$

where $\langle k\rangle(1-F(t))$ is the average susceptible neighbors to be infected for each already infected user and $\langle k\rangle$ is the average degree. Hence, for each infected user, he/she tries to induce each of his/her susceptible neighbors with probability $p_n(t) = \int_t^{t+h}\frac{\beta N}{\langle k\rangle t^\theta}dt = \frac{\beta N}{\langle k\rangle t^\theta} * h + o(h)$ during the time interval $(t, t+h)$ when $h \to 0$. As for the NETTIDE-Link, under the uniform mixing assumption that $\gamma = 1$ and α is the linear sparsity of a random graph, we decompose the aggregate hazard rate into the micro-level pairwise link building rate:

$$\begin{aligned}\frac{de(t)}{dt} &= \lambda_e(t)\times(\alpha n(t)(n(t)-1) - e(t)) \\ &= \frac{\beta'}{t^\theta}n(t)(\alpha(n(t)-1) - \frac{e(t)}{n(t)})\end{aligned} \tag{17}$$

Hence, for each infected user, he/she tries to build link with each of his/her already infected neighbors with probability $p_e(t) = \int_t^{t+h}\frac{\beta'}{t^\theta}dt = \frac{\beta'}{t^\theta} * h + o(h)$ during the time interval $(t, t+h)$ when $h \to 0$.

Although closed-from relationship between macro-level hazard rate and network-based micro-interaction are analyzed in random graph, the NETTIDE-Process can be applied to arbitrary networks. To begin with, we need the underlying organizational structure G_0, the maximal growth rate of nodes β, the temporal fizzling exponent θ, and the maximal linking rate β'. Consider the $G_1(t) = (Node(t), Edge(t))$ is the evolving network over G_0. $Node(t)$ and $Edge(t)$ are the existing nodes and links in the system G_1 at time t. We can initialize $Node(t_0)$ by random or just give the initial state as input to describe the burn-in period of the system. The same goes with $Edge(t_0)$. Thus, the NETTIDE-Process goes as follows:

- *Node growth.* For any existing node i in $Node(t)$ at time t, i tries to activate each of his neighbors, like j in G_0. If j has not existed in G_1 yet, then i tries to invite j to join with probability $p_n(t)$ during the time interval $(t, t+h)$. If success, we add (j, t) to the node set $Node(t)$, and $(i, j, t), (j, i, t)$ to the edge set $Edge(t)$.
- *Link growth.* If j has been in G_1 but not being linked to i in the G_1 yet , then i tries to build a link to j with probability $p_e(t)$ during the time interval $(t, t+h)$. If success, we add (i, j, t) and (j, i, t) to the $Edge(t)$ with timestamp t.
- *Activity.* If j has being in G_1 and being linked to i in the system already, then i can talk with (any activities supported in this specific organizational context) j, but no change to the $G_1(t)$ we care about. As the process continues, the network G_1 grows with time.

These two stochastic generators are designed to describe the stochastic growth dynamics of nodes and links. For reproducibility, we open our code, see Section 5.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2018.2801844, IEEE Transactions on Knowledge and Data Engineering

3.5 Parameter learning

The NETTIDE for node and link together has a parsimonious set of parameters, namely, $\Theta = \{\beta, \theta, \beta', \alpha, \gamma, N\}$. Our parameter learning process has two steps: to learn node equation, and to learn link equation. Given the real node growth sequence $n(t)$, we aim to minimize the sum of the square errors:

$$\underset{\beta,\theta,N}{min}\, J(n(t), n^*(t)) = \sum_{t=t_0}^{T} (n(t) - n^*(t))^2 \tag{18}$$

As for link equation, given the real link and node growth sequence $e(t)$ and $n(t)$, and the temporal fizzling exponent θ learned by the node step, we follow the same procedure as the node step to minimize the sum of the square errors:

$$\underset{\beta',\alpha,\gamma}{min}\, J(e(t), e^*(t)) = \sum_{t=t_0}^{T} (e(t) - e^*(t))^2. \tag{19}$$

We adopt the *Levenberg-Marquardt* algorithm (LM) [31] to solve these non-linear least square problems.

4 EXPERIMENTS

In this section, we evaluate the effectiveness of NETTIDE on a range of real-world evolving social networks. Here we report experiments to answer the following questions:

Q1. **Accuracy.** Can the NETTIDE capture the growth dynamics of both node and link in real-world social networks accurately?

Q2. **Usefulness.** How well do the NETTIDE forecast $n(t)$ and $e(t)$ in both near and far future?

Q3. **Generators.** Can the NETTIDE-Survival and NETTIDE-Process generate realistic growth dynamics?

4.1 Datasets

WeChat on-line social network. WeChat is the largest on-line social network in China with more than 806 million monthly active users by June 30, 2016. We collected the history data of WeChat which consists of complete records of the node and link growth from January 21, 2011 (the day WeChat was released), to January 16, 2013 when the registered users reached 300 million. In total, there are 300 million nodes (registered users rather than monthly active users) and more than 4.75 billion links. The records document the adding time of each user and the establishment timestamp of each social link. Thus, we recover the growth dynamics of both nodes and links from the inception of WeChat. We treat the bidirectional relationships between users as two links. Besides, we validate the forecasting capability of our model by five latest snapshots of the WeChat social networks from December 17, 2015 to January 14, 2016. All the WeChat data that we could access were anonymized for strict privacy policy.

ArXiv co-authorship network. This is a scientific collaboration network covering almost a decade since its inception [1]. If any two persons were in the author lists of one paper, then they formed an bidirectional link with timestamp being the date of its publication. The join date of a person is represented by the date of his first publication in this dataset. The dataset covers the period from March 1992 (near the inception of the arXiv) to March 2002. By filtering the links without explicit date, there are totally $16,959$ nodes and $2,388,880$ links.

Enron enterprise social network. Through the email records of Enron [19], we recover the enterprise social network emitted from the staff of Enron. The dataset covers the period from January 1998 to July 2002, during which Enron bankrupted on December 2, 2001, causing a sharp cut-off of the $n(t)$. In all, there are $86,458$ nodes and $594,998$ links.

Weibo information cascading network. We choose one large information cascading social network in Tencent Weibo [41], which is formed by the diffusion of a meme about a popular game. There are $165,147$ nodes and $331,607$ links, revealing the social network driven by users' interest in this game.

4.2 Q1: Accuracy

We validate the NETTIDE by answering Q1, to find out whether our model can capture the growth dynamics of node and link in real social networks.

4.2.1 Evaluation methods

We conduct the experiments in four different real social networks and set five checkpoints to give the empirical evidence for the validity and the generality of our NETTIDE model. The five checkpoints are node cumulative dynamics $n(t)$, node rate dynamics $\frac{dn(t)}{dt}$, link cumulative dynamics $e(t)$, link rate dynamics $\frac{de(t)}{dt}$, and the densification of the links against the nodes $e(n(t))$. We also consider other four methods discussed in Section 2 as baselines for comparison: Susceptible-Infected (SI), Bass model, SpikeM, and Phoenix-R (PHR). All these methods are designed for nodes, thus not applicable to links.

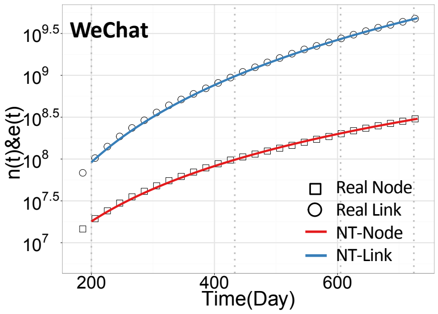
We evaluate the overall fitting accuracy by the Normalized Root Mean Square Error (*NRMSE*). Given two series, for example the real node growth sequence $n(t)$ and the corresponding sequence $n^*(t)$ given by our model, *NRMSE*$=\frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(n(t)-n^*(t))^2}}{max(n(t))-min(n(t))}$. As a special case when $T = 1$, *NRMSE* degenerates to Absolute Percentage Error (*APE*$(x, x^*) = \frac{|x-x^*|}{x}$). *NRMSE* is consistent with the objective function of the LM algorithm in the sense of $L2$ norm. And also it can be compared between datasets with different scales. We also compare the performance by other standard metric, namely Mean Absolute Percentage Error (*MAPE*). We get consistent conclusion and thus we do not report it for brevity. Table 4 shows the description of the best fitting parameters for the datasets.

4.2.2 Value and shape accuracy

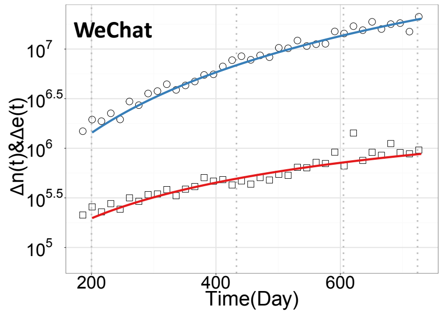
Our NETTIDE accurately fits growth dynamics of both nodes and links of WeChat, which span 726 days since the release of WeChat. The fitting results of the five checkpoints, as depicted in Fig. 3a-c, show that the growth curves generated by our NETTIDE almost overlap all the real data points. The fitting covers the period during which WeChat gained its major population: For the cumulative number, the overall errors between NETTIDE and real data are less than 1-percent, 0.76% and 0.66% for $n(t)$ and $e(t)$ respectively (Table 3). Though the rates exhibits much larger fluctuation compared with the cumulative numbers, our NETTIDE model still fits the rates well (as shown in Fig. 3b), with lowest error compared with baselines (Table 3). The densification relationship between $n(t)$ and $e(t)$ is perfectly described by NETTIDE, with overall error 1.08%. Besides, only our NETTIDE-Link is capable of capturing the link dynamics (Fig. 4b).

We then validate NETTIDE by arXiv and Enron. Our NETTIDE fits their growth dynamics accurately again, despite the facts of longer time span (5 and 10 years respectively), the tendency

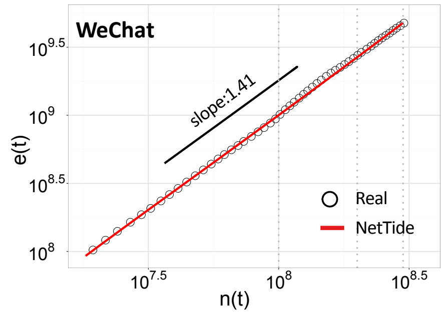
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2018.2801844, IEEE Transactions on Knowledge and Data Engineering



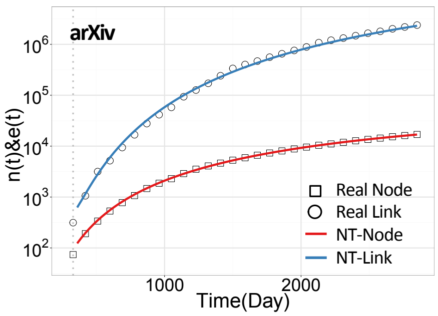
(a) WeChat cumulative



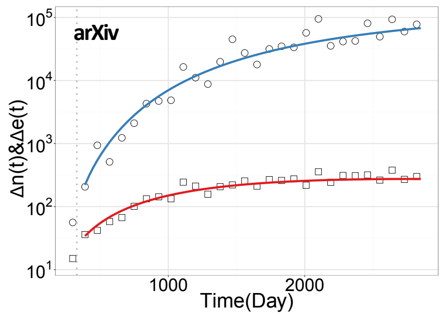
(b) WeChat rate



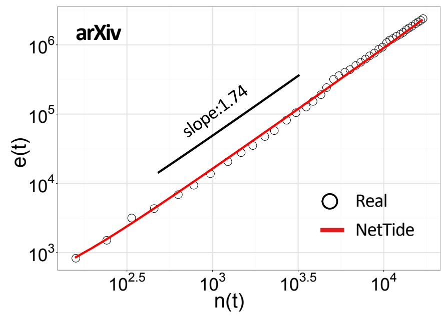
(c) WeChat densification



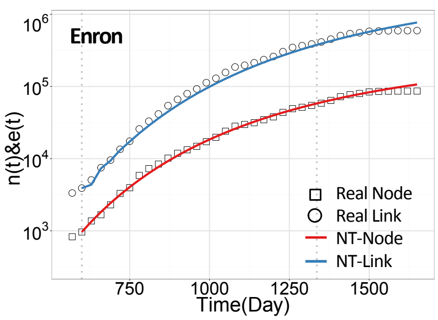
(d) arXiv cumulative



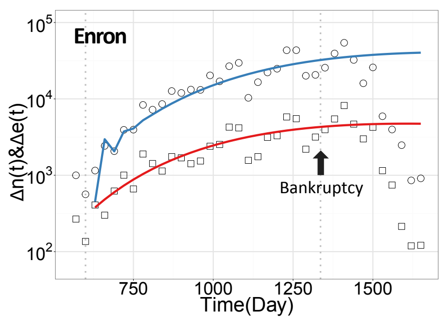
(e) arXiv rate



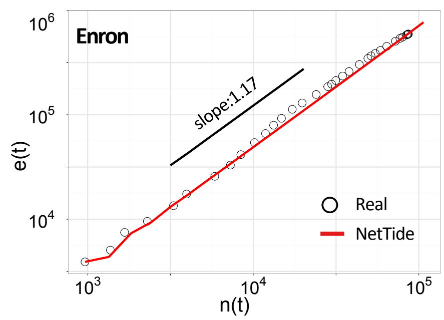
(f) arXiv densification



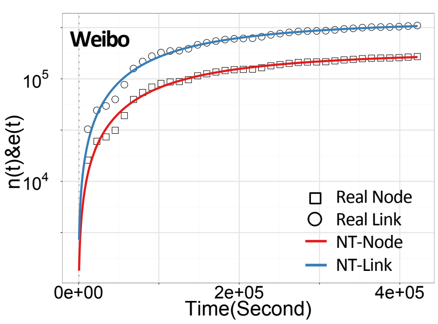
(g) Enron cumulative



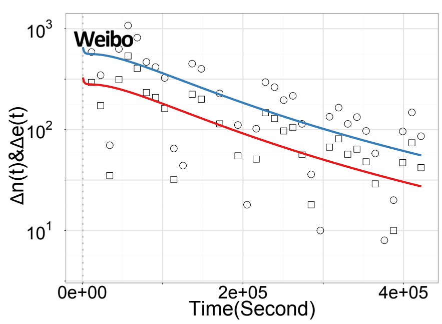
(h) Enron rate



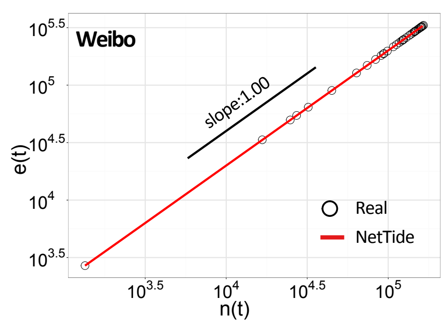
(i) Enron densification



(j) Weibo cumulative



(k) Weibo rate

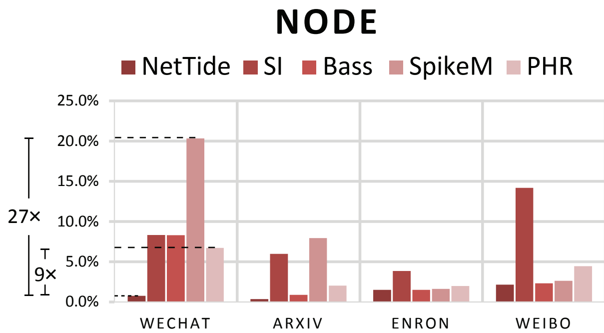


(l) Weibo densification

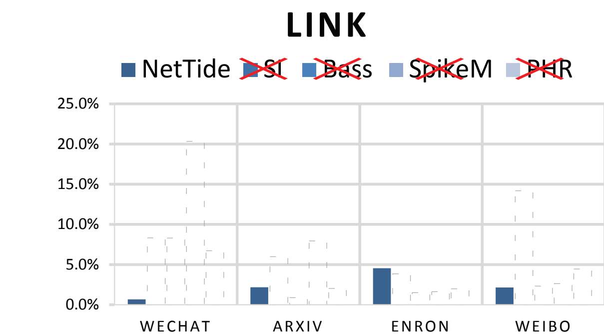
Fig. 3. NETTIDE *fits reality well.* Our model fits the growth dynamics of four real-world social networks accurately. The four rows corresponds to WeChat (a-c), arXiv (d-f), Enron (g-i) and Weibo (j-l) respectively. In each row, there are five checkpoints: $n(t)$ and $e(t)$ in the first figure, $\frac{dn(t)}{dt}$ and $\frac{de(t)}{dt}$ in the second figure, and $e(n(t))$ in the third figure.

to saturation, and the unanticipated factors (the bankruptcy of Enron). The fitting covers the period during which two social networks gained is 99-percent population. We binned the growth dynamics of these two data by month (a proper granularity for co-authorship or enterprise context). The red and blue curves by our NETTIDE almost overlap all the real data points of arXiv (Fig. 3d-f) and Enron (Fig. 3g-i). Specifically, NETTIDE-Node gets the lowest error 0.35% (1.51%) in the arXiv (Enron) case compared with baselines, as shown in Fig. 4a. Besides, NETTIDE-Link captures the link growth accurately, 2.18% and 4.54% for arXiv and Enron respectively. All the baselines are unable to describe the link growth dynamics as shown in Fig. 4b.

At last, we validate NETTIDE by Weibo, which is a volatile network and exhibits large fluctuations. Nevertheless, our NETTIDE captures the growth dynamics of Weibo well again. We binned the growth dynamics by 5 minutes because of its volatile nature. Though the daily fluctuations (ebbs and peaks corresponding to the midnight and office hour respectively as shown in



(a) Error of $n(t)$



(b) Error of $e(t)$

Fig. 4. NETTIDE *outperforms baselines*. NETTIDE-Node consistently outperform all the baselines with lowest error with respect to *NRMSE*. NETTIDE-Link fits all the datasets with very low error. All the baselines are not applicable to the links.

Fig. 3k) introduce a relatively large error (Table 3), we can still see the trends given by NETTIDE from the mess in Fig. 3j-l, and the fitting results of $n(t)$ and $e(t)$ are still good. Specifically, NETTIDE-Node and NETTIDE-Link get 2.15% and 2.15% error for $n(t)$ and $e(t)$ respectively. Still the lowest error for $n(t)$ and no baselines for $e(t)$ are shown in Fig. 4b.

Only our NETTIDE can capture the growth dynamics accurately in both value and shape aspects. So far, NETTIDE has manifested its ability to capture growth dynamics by the right shape of curves among the real points and the lowest overall fitting error. What's more, our NETTIDE-Link is unique in capturing the link growth dynamics. Thus, the rhetorical question is whether our NETTIDE-Node is also unique in its ability to capture the node growth?

All the state-of-the-art baselines fail to capture the growth of nodes in either shape or value aspects: The exponential growth nature of SI and Bass at early stage deviates from the real data seriously, causing failure in modeling the growth of WeChat with power law growth. The SI and Bass have very similar performances in the WeChat case, with errors up to 10.0 times greater than the results of NETTIDE-Node. In other datasets, SI also deviates from the real seriously as shown in Fig. 4a, with errors 16.1, 1.5 and 5.6 times greater than our fitting for arXiv, Enron and Weibo respectively. Though the incorporating of market growth in Bass model reduces the error compared with SI, the exponential shape of Bass curve at early stage is totally wrong with our power-law like observations. The performance of SpikeM in different datasets varies a lot. The best fitting of SpikeM in the WeChat and Weibo cases lie in the sub-critical

TABLE 3
The comparison of the accuracy of NETTIDE and baseline methods on five checkpoints of the growth dynamics of four real social networks. The NETTIDE consistently outperforms baselines in almost every check point. All the baselines are not applicable (—) to the links.

WeChat	**n(t)**	**e(t)**	**dn(t)/dt**	**de(t)/dt**	**e(n)**
NETTIDE	**0.76%**	**0.66%**	**6.29%**	**5.07%**	**1.08%**
SI	8.32%	—	23.39%	—	—
BASS	8.31%	—	23.64%	—	—
SPIKEM	20.33%	—	48.19%	—	—
PHR	6.73%	—	8.59%	—	—
arXiv	**n(t)**	**e(t)**	**dn(t)/dt**	**de(t)/dt**	**e(n)**
NETTIDE	**0.35%**	**2.18%**	**9.91%**	**11.27%**	**3.32%**
SI	5.97%	—	33.83%	—	—
BASS	0.88%	—	11.18%	—	—
SPIKEM	7.95%	—	24.63%	—	—
PHR	2.03%	—	15.07%	—	—
Enron	**n(t)**	**e(t)**	**dn(t)/dt**	**de(t)/dt**	**e(n)**
NETTIDE	**1.51%**	**4.54%**	14.62%	**14.27%**	**4.62%**
SI	3.84%	—	20.74%	—	—
BASS	1.51%	—	**14.54%**	—	—
SPIKEM	1.63%	—	18.00%	—	—
PHR	1.99%	—	15.89%	—	—
Weibo	**n(t)**	**e(t)**	**dn(t)/dt**	**de(t)/dt**	**e(n)**
NETTIDE	**2.15%**	**2.15%**	14.93%	**14.90%**	**0.06%**
SI	14.19%	—	24.51%	—	—
BASS	2.31%	—	15.01%	—	—
SPIKEM	2.62%	—	**14.78%**	—	—
PHR	4.45%	—	17.53%	—	—

regime of the hawkes process. However, the SpikeM reports the largest errors (25.8 times greater than NETTIDE-Node) in the WeChat case, while a relatively low error (21.9% greater than NETTIDE-Node) is reached in Weibo. The super-critical regime, which generates exponential growth at early stage, is reached in fitting the arXiv and Enron, with errors 21.7 times and 8.0% greater than NETTIDE-Node respectively. The problems of wrong shape and largely fluctuated errors also come with Phoenix-R: it reports the lowest error among the baselines in WeChat, still 7.9 times larger than our NETTIDE-Node. The errors of Phoenix-R are 4.8, 1.1 times greater than NETTIDE-Node for for arXiv and Weibo, and 31.8% greater than NETTIDE-Node for Enron.

In all, only our NETTIDE correctly approximates the node and link growth dynamics of real social networks, in both value and shape aspects.

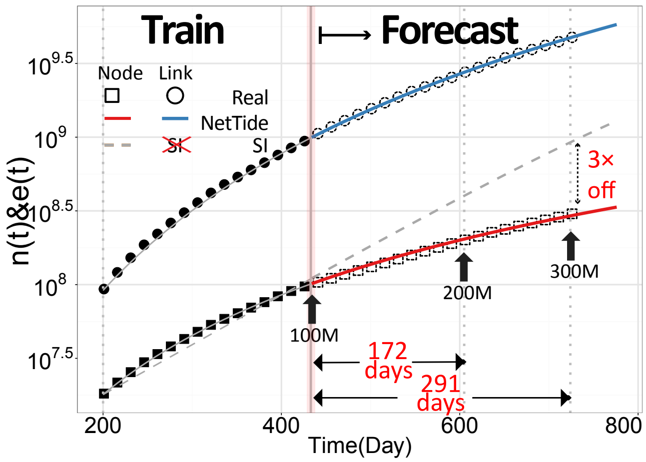
4.2.3 Parameter analysis

TABLE 4
The parameters of NETTIDE best fitting each dataset.

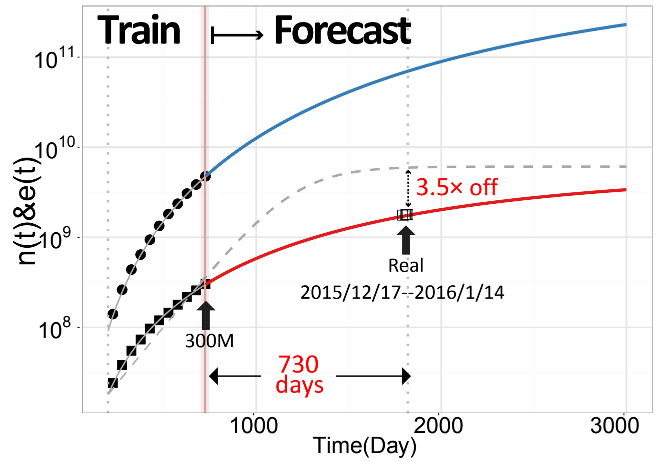
	N	βN	θ	β'	α	γ
WeChat	6.1B	2.16	0.995	0.03	0.14	0.47
arXiv	12584	8.81	1.35	7.56	0.28	0.74
Enron	458143	155.14	1.96	751.19	1.30	0.16
Weibo	18935	0.50	0.84	0.030	1.68	0.02

The best fitting parameters of NETTIDE accurately meet the characteristics of real growth dynamics of each social network (as shown in Table 4). In WeChat case, the value of fizzling exponent θ (0.995 by NETTIDE) is very close to 1, implying the power law growth of the nodes, with exponent $\approx \beta N = 2.16$ (close

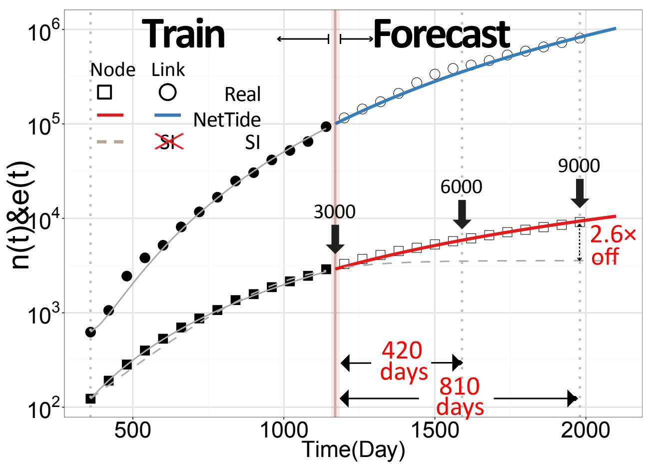
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2018.2801844, IEEE Transactions on Knowledge and Data Engineering



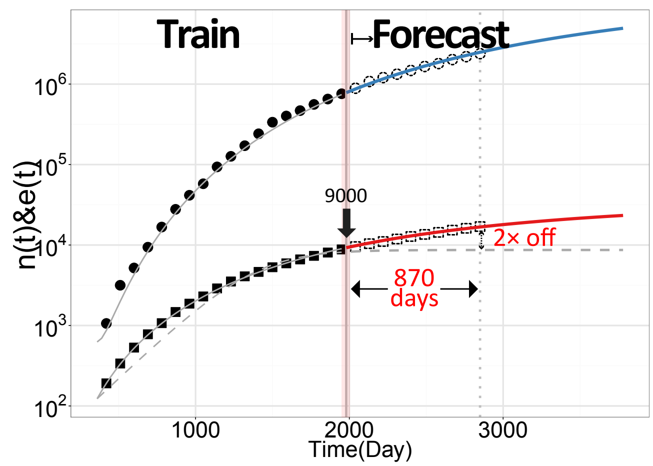
(a) Short-term forecasting for WeChat



(b) Long-term forecasting for WeChat



(c) Short-term forecasting for arXiv



(d) Long-term forecasting for arXiv

Fig. 5. NETTIDE *forecasts future well.* The points represent real data, black filled for training, and the dashed for validation. The red and blue lines are the forecasting results of NETTIDE-Node and NETTIDE-Link respectively. The gray dashed lines are the results of SI. The above panel is the results of WeChat, while the panel below is the results of arXiv. (a) and (c) are the results of the short term forecasting, while (b) and (d) are the results of the long term case.

to 2.15 of the real $n(t)$, in Figure 1a). The exponent of the densification power law is revealed by $1+\gamma = 1.47$ (close to 1.41 of the real $e(n(t))$ as shown in Fig 3c). As for the arXiv, the learned $\theta = 1.35$, implies a relatively faster temporal fizzling effect than the Log-Logistic curve with $\theta = 1$. The evolution of arXiv network has the largest exponent of densification power law 1.74 among four datasets, captured by the $\gamma = 0.74$, implying the tightly-knit structure of the high energy physics community. The fizzling exponent (1.96) of Enron is the largest among the datasets, capturing the quick saturation to the ceiling caused by the bankruptcy of Enron. The structure of the company of Enron shows the hierarchical structure like tree, due to the relatively small exponent of the densification power law 1.17 (captured by NETTIDE $\gamma = 0.16$), when compared with other social network data. The tree structure of WeChat shows linear growth of link with node, which makes the γ ($\gamma = 0.02$ by NETTIDE) very close to zero. The best fitting value $\theta = 0.84$, exhibits a faster-than power-law growth at early stage due to the fast growth nature of Weibo.

4.3 Q2: Usefulness-forecasting

We show the practical value of our NETTIDE by answering Q2, to forecast both the count of nodes and links, in the short term and in the long term.

4.3.1 *Short-term forecasting*

In the short-term forecasting setting, we validate NETTIDE's forecasting capability by examining the overall predictive error into the future (overall forecasting task) and the arrival of some checkpoints marked as milestones (milestone forecasting task). Specifically, taking WeChat as an example, by training the dynamics of nodes within first 100 million : the overall forecasting task is to examine how well NETTIDE-Node forecast the growth dynamics of next 200 million nodes; the milestone forecasting task is to forecast the date when WeChat network doubles and triples its size. We denote the t_1, t_2, t_3 as the date of the milestones. In WeChat case, they are the dates when WeChat network hit its first 100, 200, 300 million nodes respectively, as shown in Fig. 5a. In arXiv case, they are the dates of reaching 3000, 6000, 9000 authors respectively, in Fig. 5c. The same task goes with NETTIDE-Link, in which case the number of links is never predicted before.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2018.2801844, IEEE Transactions on Knowledge and Data Engineering

Overall forecasting. Both NETTIDE-Node and NETTIDE-Link can forecast future dynamics very accurately, covering 291 and 730 days in the future for WeChat and arXiv respectively. In the WeChat case, the overall errors are 2.18% for $n(t)$ and 0.44% for $e(t)$ between the forecasting results by NETTIDE and the real dynamics from t_1 to t_3 (Fig 5a). For the arXiv, the overall errors are 2.86% and 4.18% for $n(t)$ and $e(t)$ respectively (Fig. 5c) from t_1 to t_3 . We also compare the forecasting results by SI: the sigmoid curve seriously overestimates the growth with overall error 134.62% for $n(t)$ in WeChat case, and underestimates the $n(t)$ of arXiv with overall error 52.14%. The SI is not applicable to the $e(t)$ (no dashed lines for Link in Fig. 5).

Milestone forecasting. Both NETTIDE-Node and NETTIDE-Link can forecast the arrival of milestones with very low error, both for the date and the count. Specifically, in WeChat case shown in Fig 5a, NETTIDE-Node forecast the arrival of first 200 million nodes 5 days earlier than the real date t_2 (172 days ahead into the future), and the arrival of first 300 million nodes 10 days later then real date t_3 (291 days ahead into the future). At t_2 and t_3, the forecasting errors are 1.67% and 2.58% for $n(t)$, and 0.26% and 0.33% for $e(t)$ respectively. As for the arXiv network, despite the fact that the t_2 (t_3) is 420 (810) days ahead into the future, NETTIDE-Node can forecast the arrival of the milestones (6000, 9000 authors) within one month centering the real date. (The time granularity we choose is just one month for arXiv and Enron.) The forecasting errors at t_2 and t_3 are 0.91% and 2.47% for $n(t)$ respectively, while 11.32% and 2.75% for $e(t)$. In contrast, the results of nodes predicted by SI are seriously biased: in WeChat case, 93 days earlier for t_2 and 167 days earlier for t_3; the deviations increase with time, more than 300% deviation at t_3. As for the arXiv, SI seriously underestimates the number of the nodes at milestones: more than 260% underestimation at t_3. Again, there are no baselines for link growth.

In all, our NETTIDE achieves a surprisingly high forecasting accuracy for both node and link growth in the short term.

4.3.2 Long-term forecasting, 2 years ahead

Our NETTIDE also shows accurate forecasting results in the long term, 730 and 870 days ahead into the future for WeChat and arXiv respectively.

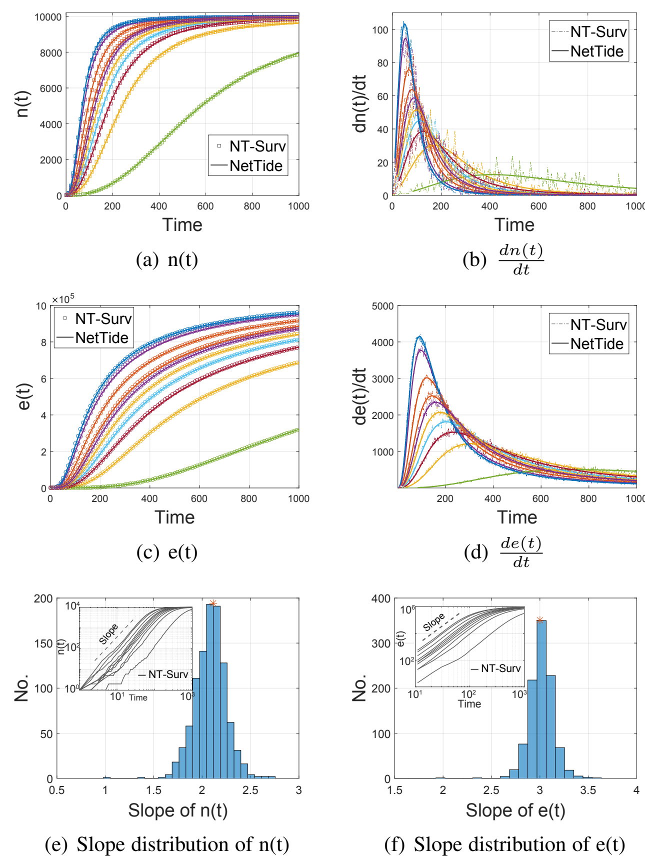
As for the WeChat case, NETTIDE-Node can forecast the number of nodes 730 days ahead into the future accurately (Fig 5b). We train NETTIDE-Node by the growth dynamics before t_3, and then we validate the forecasting results of NETTIDE in the long term by 5 latest snapshots of the WeChat social network. The 5 latest checkpoints span more than one month (December 17, 25, 2015, and January 1, 8, 14, 2016). For the privacy issues, we do not report the exact number of registered users and the number of links. We set the initial total population N to be 6.1 billion, the smart-phone users globally by 2020, reported by Ericsson[2]. Because one user can only register the WeChat successfully through the verification of his phone number. The errors for $n(t)$ at these five checkpoints are consistently low, 2.86%, 2.72%, 2.68%, 2.68% and 2.64% for each checkpoints respectively. However, the node growth curve of SI seriously overestimates the real node growth: the saturation point is reached much earlier, and with 350% deviation with the real data at 2016/1/14.

In the arXiv case, NETTIDE can forecast both the $n(t)$ and $e(t)$ accurately in the long term, 870 days ahead into the future as shown in Fig 5d. We train both NETTIDE-Node and NETTIDE-Link by the real growth dynamics before t_3, and we get overall error 2.84% for $n(t)$ and 3.56% for $e(t)$, covering 870 days in the future. However, the forecasting results of the number of nodes by SI seriously underestimates the real number, up to 200% off the reality.

2. http://www.ericsson.com/mobility-report

4.4 Q3: Generators

So far, we have examined the accuracy and usefulness of our NETTIDE model. Here we generate realistic growth dynamics as WeChat by both two generators, i.e., the NETTIDE-Survival generator and the NETTIDE-Process generator.



(a) n(t) (b) $\frac{dn(t)}{dt}$

(c) e(t) (d) $\frac{de(t)}{dt}$

(e) Slope distribution of n(t) (f) Slope distribution of e(t)

Fig. 6. NETTIDE-Survival *generates realistic stochastic dynamics.* (a-b) Ten stochastic trajectories of n(t) (squares □) and $\frac{dn(t)}{dt}$ (dash-dot line) generated from NETTIDE-Survival respectively. The solid lines are the fitting results by our NETTIDE model. Each specific color corresponds to one generated instance. (c-d) are the counterparts for links. (e-f) The histograms of power-law exponents for generated n(t) and e(t) respectively. $1,000$ n(t) and e(t) trajectories are generated by NETTIDE-Survival. The red asterisks denote the real exponents in WeChat case. The curves in inset on log-log scale are the same n(t) and e(t) as in (a-b), which exhibit power-law early growth.

4.4.1 Stochastic dynamics from NETTIDE-Survival

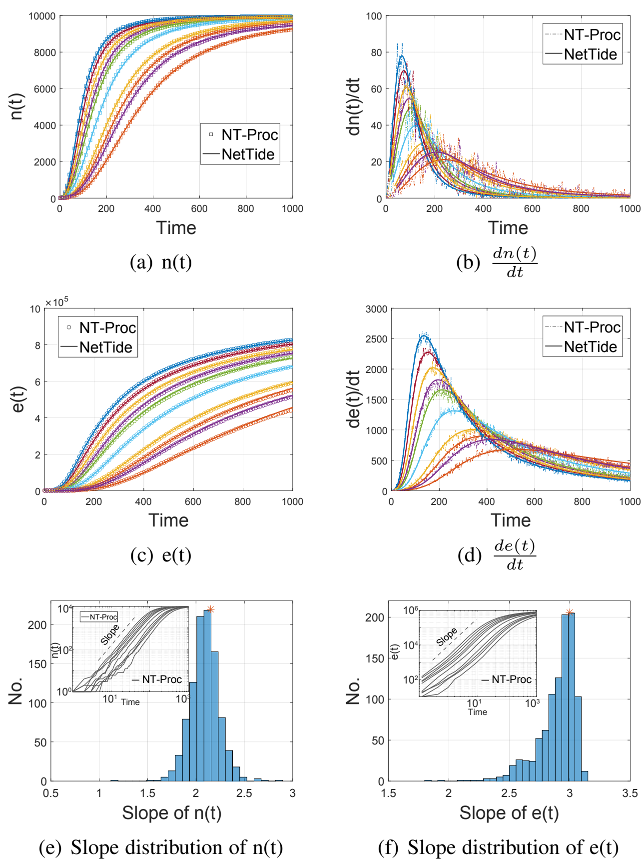
The NETTIDE-Survival generator generates stochastic trajectories of node and link growth from hazard rates directly as shown in Sec3.4. For different combination of modeling parameters $(\beta, \theta, N, \beta', \alpha, \gamma)$, NETTIDE-Survival can generate a wide range of node and link trajectories. Here we report one specific setting $(\beta = 2.45 \times 10^{-4}, \theta = 1, N = 10^4, \beta' = 0.5, \alpha = 0.9, \gamma = 0.52)$ which generates similar growth dynamics as WeChat, which

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2018.2801844, IEEE Transactions on Knowledge and Data Engineering

follows power-law growth with exponents 2.15 and 3.01 for nodes and links respectively. We open our code for generating stochastic growth dynamics in arbitrary parameter settings.

Figures 6a-d illustrate ten stochastic instances generated by the NETTIDE-Survival which are distinguished by different colors. Although the NETTIDE-Survival generates quite different $n(t)$, $\frac{dn(t)}{dt}$, $e(t)$, and $\frac{de(t)}{dt}$ with stochastic fluctuations, surprisingly, our NETTIDE capture all the stochastic trajectories quite well. Due to fact that $\theta = 1$, the NETTIDE-Survival generate power-law-like growth dynamics, as shown in insets of Figs. 6e-f. We generate $1,000$ n(t) and e(t) instances by NETTIDE-Survival and fit the generated power-law early growth of by least square method. Figures 6c&f show the distributions of the power-law exponents of node and link growth respectively, indicating that the NETTIDE-Survival generate n(t) and e(t) with power-law exponents near 2.15 and 3, implying that the NETTIDE-Survival successfully reproduces the WeChat growth dynamics.

4.4.2 Stochastic dynamics from NETTIDE-Process



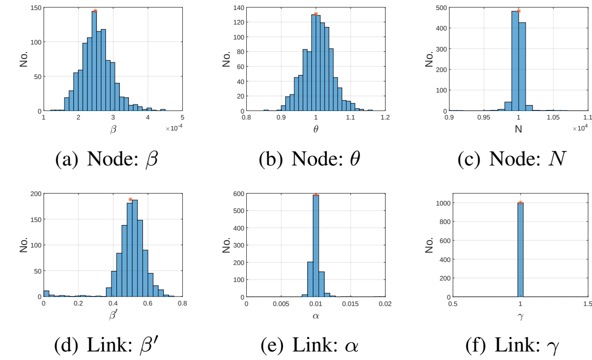
(a) n(t) (b) $\frac{dn(t)}{dt}$

(c) e(t) (d) $\frac{de(t)}{dt}$

(e) Slope distribution of n(t) (f) Slope distribution of e(t)

Fig. 7. NETTIDE-Process *generates realistic stochastic dynamics.* (a-b) Ten stochastic trajectories of n(t) (squares □) and $\frac{dn(t)}{dt}$ (dash-dot line) generated from NETTIDE-Process respectively. The solid lines are the fitting results of the generated stochastic trajectories by our NETTIDE model. Each specific color corresponds to one generated instance. (e) The histogram of power-law exponents of $1,000$ n(t) trajectories generated from NETTIDE-Process. The red asterisk denotes the real exponent in WeChat case. The curves in inset on log-log scale are the same n(t) as in (a), which exhibit power-law early growth. (cd&f) are the counterparts for links.

We further analyze the second stochastic generator of the NETTIDE from micro-level stochastic interactions within a network, i.e., the NETTIDE-Process. Given the underlying organizational



(a) Node: β (b) Node: θ (c) Node: N

(d) Link: β' (e) Link: α (f) Link: γ

Fig. 8. NETTIDE-Process *can be inferred by* NETTIDE. The parameters of the NETTIDE-Process and the underlying random graph are inferred by our NETTIDE model. (a-f) The histograms of the inferred parameters of the same $1,000$ instances as in Fig. 6. The asterisks in red denote the real values of the parameters used in NETTIDE-Survival. In the reference process, we set $\gamma = 1$ for the random graph case.

structure G_0, the modeling parameters we need for simulation are (β, θ, β'). As discussed in model section, under the random graph assumption, the micro-pairwise infection probability $p_n(t)$ and linking probability $p_e(t)$ during time interval $(t, t+h)$ can be derived from equation 16 and 17 respectively. We again report our results under the setting $(\beta = 2.485 \times 10^{-4}, \theta = 1, \beta' = 0.48)$ and $G_0 = RandomGraph(\alpha = 0.01, N = 10^4)$, which again generates similar growth dynamics as WeChat case. For generating stochastic trajectories in arbitrary parameter settings can be realized by our open-sourced code.

The stochastic growth dynamics generated by NETTIDE-Process are shown in Fig. 7. Again, the solid curves almost hit every dot, indicating that NETTIDE fits the stochastic $n(t)$, $\frac{dn(t)}{dt}$, $e(t)$, and $\frac{de(t)}{dt}$ well. By using the NETTIDE-Process with aforementioned parameters, we generate $1,000$ stochastic growth instances. We find the generated trajectories show power-law early growths (illustrated in the insets of Figs. 7e-f), and the power-law exponents of n(t) and e(t) are close to 2.15 and 3, implying the NETTIDE-Process generate realistic growth dynamics as WeChat case.

We further examine whether NETTIDE can uncover the "real" parameters which NETTIDE-Process used. In fitting process, we set the $\gamma = 1$ due to the random graph setting. We find NETTIDE infers the modeling parameters as shown in Fig. 8. Our NETTIDE not only finds the values of β, θ and β' which controls behaviors, but also uncovers the structure parameters $N = 10^4$ and $\alpha = 0.01$.

In all, both NETTIDE-Survival and NETTIDE-Process can generate realistic growth dynamics of the node and link, and the generated stochastic growth dynamics can be well captured by NETTIDE well.

5 CONCLUSIONS

In this paper, we studied the growth dynamics of real-world social networks and presented NETTIDE to capture growth dynamics of both nodes and links. We examine a range of real-world evolving social networks, especially China's largest online social network WeChat. We find both node and link in real social networks follow Power-Law growth, rather than the exponential growth or uniform growth as expected. Thus, we propose NETTIDE, along with differential equations for the growth of the number of nodes, as well as links. Our NETTIDE-Node gives a unified but parsimonious model to capture real social network growth, like the power law early growth of the Log-Logistic growth, and more general form Stretched-Exponential early growth of the Stretched-Logistic growth. Our NETTIDE-Link is the first-ever differential equation to capture the growth dynamics of links, accurately fitting reality. Furthermore, we propose two stochastic generators, i.e., NETTIDE-Survival and NETTIDE-Process, which generate realistic growth dynamics from the perspective of survival analysis and microscopic level stochastic interactions within a network respectively. Our NETTIDE again accurately fits the stochastic generators, and infers the parameters of the generators well. The main contributions are:

1) **Novel model NETTIDE**: NETTIDE-Node captures a wide range of growth dynamics and NETTIDE-Link is the first differential equation to capture the link growth dynamics. Both equations are parsimonious and explainable on microscopic level.
2) **Accuracy**: We presented experiments on four real-world evolving social networks, especially the WeChat (300 million nodes, 4.75 billion links). Our NETTIDE model matches the real-world growth dynamics accurately.
3) **Usefulness**: Our NETTIDE can be used to both the short-term and long-term forecasting. We validated NETTIDE's forecast power empirically, and showed that it can forecast the nodes and links in the short term and even the long term accurately (730 and 870 days ahead into the future for WeChat and arXiv respectively).
4) **Generators**: We propose two stochastic generators at microscopic level, i.e., NETTIDE-Survival and NETTIDE-Process, which successfully generate realistic stochastic growth dynamics.

Reproducibility: We have already open-sourced our code of the NETTIDE together with two generators NETTIDE-Survival and NETTIDE-Process, to fit/generate the deterministic/stochastic growth dynamics of both nodes and links, at https://github.com/calvin-zcx/NetTide

There exist many directions of further studies. First, the proposed growth models can be applied to validate the ubiquitous growth phenomena in other fields, like the ecology, social science, demography, and so on. Second, inspired by the physical meanings of the modeling parameters of NETTIDE, how to improve the social network services and boost the social behaviors are open questions. Third, one major limitation of our NETTIDE is the neglect of external influences. How the external signals influence the growth or decay dynamics of social networks remains to be examined.

Acknowledgments

This work was supported in part by National Program on Key Basic Research Project (No. 2015CB352300), National Natural Science Foundation of China (No. U1611461, No. 61772304, No. 61521002, No. 61531006), the research fund of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology, and the Young Elite Scientist Sponsorship Program by CAST.

References

[1] arxiv hep-ph network dataset – KONECT, May 2015.

[2] R. Albert and A.-L. Barabási. Topology of evolving networks: local events and universality. *Physical review letters*, 85(24):5234, 2000.

[3] R. M. Anderson, R. M. May, and B. Anderson. *Infectious diseases of humans: dynamics and control*, volume 28. Wiley Online Library, 1992.

[4] D. Antoniades and C. Dovrolis. Co-evolutionary dynamics in social networks: A case study of twitter. *Computational Social Networks*, 2(1):1–21, 2015.

[5] R. B. Banks. *Growth and diffusion phenomena: mathematical frameworks and applications*, volume 14. Springer Science & Business Media, 1994.

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

[7] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[8] A.-L. Barabâsi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3):590–614, 2002.

[9] M. Berlingerio, F. Bonchi, B. Bringmann, and A. Gionis. Mining graph evolution rules. In *Machine learning and knowledge discovery in databases*, pages 115–130. Springer, 2009.

[10] G. Bianconi and A.-L. Barabási. Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)*, 54(4):436, 2001.

[11] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

[12] M. Farajtabar, Y. Wang, M. Rodriguez, S. Li, H. Zha, and L. Song. Coevolve: A joint point process model for information diffusion and network co-evolution. In *Advances in Neural Information Processing Systems*, pages 1945–1953, 2015.

[13] F. Figueiredo, J. M. Almeida, Y. Matsubara, B. Ribeiro, and C. Faloutsos. Revisit behavior in social media: The phoenix-r model and discoveries. In *Machine Learning and Knowledge Discovery in Databases*, pages 386–401. Springer, 2014.

[14] G. Ghoshal, L. Chi, and A.-L. Barabási. Uncovering the role of elementary processes in network evolution. *Scientific reports*, 3, 2013.

[15] R. Gonzalez, R. Cuevas, R. Motamedi, R. Rejaie, and A. Cuevas. Assessing the evolution of google+ in its first two years. *Biological Cybernetics*, 24(3):1813–1826, 2016.

[16] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[17] B. A. Huberman and L. A. Adamic. Internet: growth dynamics of the world-wide web. *Nature*, 401(6749):131–131, 1999.

[18] M. Jiang, P. Cui, and C. Faloutsos. Suspicious behavior detection: current trends and future directions. *Special Issue on Online Behavioral Analysis and Modeling, IEEE Intelligent Systems Magazine*, 2016.

[19] B. Klimt and Y. Yang. The Enron corpus: A new dataset for email classification research. In *Proc. European Conf. on Machine Learning*, pages 217–226, 2004.

[20] G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.

[21] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. KDD '06, pages 611–617, New York, NY, USA, 2006. ACM.

[22] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 57–65. IEEE, 2000.

[23] P. J. Laub, T. Taimre, and P. K. Pollett. Hawkes processes. *arXiv preprint arXiv:1507.02822*, 2015.

[24] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. KDD '08, pages 462–470. ACM, 2008.

[25] J. Leskovec, D. Chakrabarti, J. Kleinberg, and C. Faloutsos. Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. In *Knowledge Discovery in Databases: PKDD 2005*, pages 133–145. Springer, 2005.

[26] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2, 2007.

[27] L. Li, B. A. Prakash, and C. Faloutsos. Parsimonious linear fingerprinting for time series. *Proceedings of the VLDB Endowment*, 3(1-2):385–396, 2010.

[28] Y. Lin, A. A. Raza, J.-Y. Lee, D. Koutra, R. Rosenfeld, and C. Faloutsos. Influence propagation: Patterns, model and a case study. In *Advances in Knowledge Discovery and Data Mining*, pages 386–397. Springer, 2014.

[29] V. Mahajan, E. Muller, and F. M. Bass. New product diffusion models in marketing: A review and directions for research. *The journal of marketing*, pages 1–26, 1990.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2018.2801844, IEEE Transactions on Knowledge and Data Engineering

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

[31] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics*, 11(2):431–441, 1963.

[32] Y. Matsubara, Y. Sakurai, and C. Faloutsos. Autoplait: Automatic mining of co-evolving time sequences. In *Proceedings of the 2014 ACM SIGMOD*, pages 193–204. ACM, 2014.

[33] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos. Rise and fall patterns of information diffusion: model and implications. KDD '12, pages 6–14. ACM, 2012.

[34] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the flickr social network. In *Proceedings of the first workshop on Online social networks*, pages 25–30. ACM, 2008.

[35] J. G. Oliveira and A.-L. Barabási. Human dynamics: Darwin and einstein correspondence patterns. *Nature*, 437(7063):1251–1251, 2005.

[36] B. Ribeiro. Modeling and predicting the growth and death of membership-based websites. In *Proceedings of the 23rd international conference on World Wide Web*, pages 653–664. ACM, 2014.

[37] H. D. Rozenfeld, D. Rybski, J. S. Andrade, M. Batty, H. E. Stanley, and H. A. Makse. Laws of population growth. *Proceedings of the National Academy of Sciences*, 105(48):18702–18707, 2008.

[38] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):036127, 2006.

[39] L. Weng, J. Ratkiewicz, N. Perra, B. Gonçalves, C. Castillo, F. Bonchi, R. Schifanella, F. Menczer, and A. Flammini. The role of information diffusion in the evolution of social networks. In *Proceedings of the 19th ACM SIGKDD'13*, pages 356–364. ACM, 2013.

[40] 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*, 107(44):18803–18808, 2010.

[41] L. Yu, P. Cui, F. Wang, C. Song, and S. Yang. From micro to macro: Uncovering and predicting information cascading process with behavioral dynamics. In *IEEE International Conference on Data Mining, ICDM*, 2015.

[42] L. Yu, P. Cui, F. Wang, C. Song, and S. Yang. Uncovering and predicting the dynamic process of information cascades with survival model. *Knowledge and Information Systems*, pages 1–27, 2016.

[43] C. Zang, P. Cui, and C. Faloutsos. Beyond sigmoids: The nettide model for social network growth, and its applications. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 2015–2024. ACM, 2016.

[44] C. Zang, P. Cui, C. Faloutsos, and W. Zhu. 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*, KDD '17, pages 565–574, New York, NY, USA, 2017. ACM.

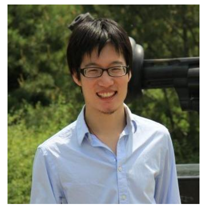
[45] C. Zang, P. Cui, C. Song, C. Faloutsos, and W. Zhu. Quantifying structural patterns of information cascades. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 867–868. International World Wide Web Conferences Steering Committee, 2017.

[46] T. Zhang, P. Cui, C. Faloutsos, Y. Lu, H. Ye, W. Zhu, and S. Yang. 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*, pages 1355–1364, 2016.

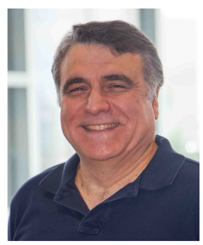
[47] T. Zhang, P. Cui, C. Song, W. Zhu, and S. Yang. A multiscale survival process for modeling human activity patterns. *PloS one*, 11(3):e0151473, 2016.

[48] K. Zhu, W. Li, X. Fu, and J. Nagler. How do online social networks grow? *PloS one*, 9(6):e100023, 2014.

Chengxi Zang received the BE degree from the Department of Computer Science and Technology of Sun Yat-sen University in 2013. He is a Ph.D. candidate in the Department of Computer Science and Technology of Tsinghua University. His main research interests include modeling structure and dynamics of large-scale social systems. He has published papers on social network growth and information cascades in top data-mining conferences, e.g., ACM SIGKDD, WWW, etc.



Peng Cui received his Ph.D. degree in computer science in 2010 from Tsinghua University and he is an Associate Professor at Tsinghua. He has vast research interests in data mining, multimedia processing, and social network analysis. Until now, he has published more than 20 papers in conferences such as SIGIR, AAAI, ICDM, etc. and journals such as IEEE TMM, IEEE TIP, DMKD, etc. Now his research is sponsored by National Science Foundation of China, Samsung, Tencent, etc. He also serves as Guest Editor, Co-Chair, PC member, and Reviewer of several high-level international conferences, workshops, and journals.



Christos Faloutsos is a Professor at Carnegie Mellon University. He has received the Presidential Young Investigator Award by the National Science Foundation (1989), the Research Contributions Award in ICDM 2006, the Innovations award in KDD10, 20 "best paper" awards, and several teaching awards. He has served as a member of the executive committee of SIGKDD; he has published over 200 refereed articles, 11 book chapters and one monograph. He holds five patents and he has given over 30 tutorials and over 10 invited distinguished lectures. His research interests include data mining for graphs and streams, fractals, database performance, and indexing for multimedia and bio-informatics data.

Wenwu Zhu received the PhD degree from Polytechnic Institute of New York University in 1996. He is currently a professor at Tsinghua University. His research interest is wireless/Internet multimedia communication and computing. He worked at Bell Labs during 1996 to 1999. He was with Microsoft Research Asia's Internet Media Group and Wireless and Networking Group as research manager from 1999 to 2004. He was the director and chief scientist at Intel Communication Technology Lab, China. He was also a senior researcher at the Internet Media Group at Microsoft Research Asia. He has published more than 200 referred papers and led 40 patents. He participated in the IETF ROHC WG on robust TCP/IP header compression over wireless links and IEEE 802.16m WG standardization. He currently serves as a chairman of IEEE Circuits and System Society Beijing Chapter, advisory board on the International Journal of Handheld Computing Research, and president of ACM Beijing. He is a fellow of the IEEE.