Beyond Sigmoids: the NetTide Model for Social Network Growth, and its Applications

Chengxi Zang[1], Peng Cui[1], Christos Faloutsos[2]
[1]Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University. Beijing, China
[2] Computer Science Department, Carnegie Mellon University, PA, USA
zangcx13@mails.tsinghua.edu.cn,cuip@tsinghua.edu.cn, christos@cs.cmu.edu

ABSTRACT

What is the growth pattern of social networks, like Facebook and WeChat? Does it truly exhibit exponential early growth, as predicted by textbook models like the Bass model, SI, or the Branching Process? How about the count of links, over time, for which there are few published models?

We examine the growth of several real networks, including one of the world's largest online social network, "WeChat", with *300 million* nodes and *4.75 billion* links by 2013; and we observe *power law* growth for both nodes and links, a fact that completely breaks the sigmoid models (like SI, and Bass). In its place, we propose NETTIDE, along with differential equations for the growth of the count of nodes, as well as links. Our model *accurately* fits the growth patterns of real graphs; it is *general*, encompassing as special cases all the known, traditional models (including Bass, SI, log-logistic growth); while still remaining *parsimonious*, requiring only a handful of parameters. Moreover, our NETTIDE for link growth is the first one of its kind, *accurately* fitting real data, and naturally leading to the densification phenomenon. We validate our model with four real, time-evolving social networks, where NETTIDE gives good fitting accuracy, and, more importantly, applied on the WeChat data, our NETTIDE model forecasted more than *730* days into the future, with *3%* error.

Keywords

Social networks; Growth model; Power law growth; Fizzle Logistic; Link growth

1. INTRODUCTION

How many members will Twitter have, next month? How many friendship links will FaceBook (or WeChat[1], 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 indicate resistance to Alzheimer's disease, etc).

Network growth models: Researchers from multiple disciplines have studied network growth phenomenon for decades [7, 24, 20, 23, 31, 19], 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 [27] 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 the following questions of social network growth:

1. How does the number of nodes $n(t)$ grow over time?
2. How does the number of links $e(t)$ grow over time?

Reality check: 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 [18] and Weibo [35], respectively representing on-line social communication networks, co-authorship 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 surprisingly 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 far beyond our traditional expectations of exponential or uniform network growth dynamics.

Our design goals: Since sigmoids and related textbook models are contradicted by reality, we need a better model. We shoot for a model that will fulfill the following GOALS:

G1. *Parsimony:* The model should have as few parameters as necessary, and still generate power-law early-growth.

G2. *All encompassing:* The model should be general, encompassing as special cases all the known, traditional models (like Bass, SI, and Log-Logistic growth).

G3. *Link growth:* The model should be able to accurately capture the growth dynamics of links.

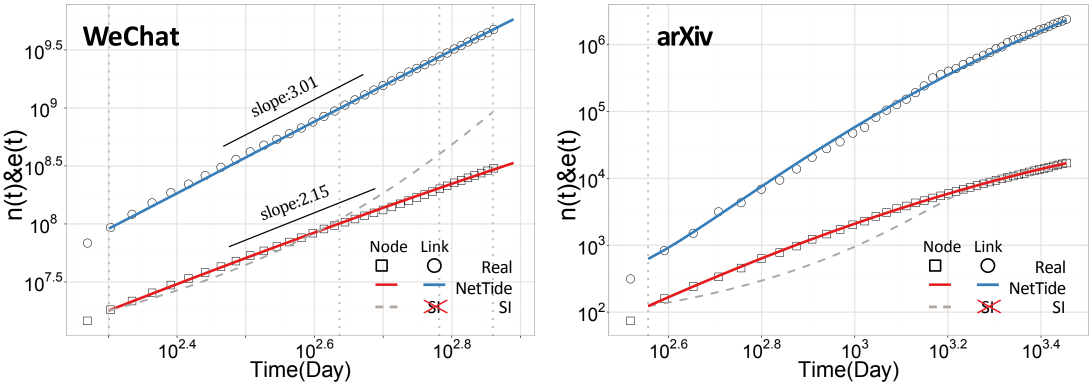
G4. *Intuition:* The model should easy to explain, with intuitive arguments.

We propose a novel dynamic model, named NETTIDE, for net-

[1]www.wechat.com/en/

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '16, August 13-17, 2016, San Francisco, CA, USA
© 2016 ACM. ISBN 978-1-4503-4232-2/16/08...$15.00
DOI: http://dx.doi.org/10.1145/2939672.2939825



(a) WeChat

(b) arXiv

Figure 1: *Reality disobeys sigmoids*: WeChat (a), and arXiv (b) - nodes over time (squares □), well fitted by our proposed NETTIDE-Node (solid red line), but not by the SI 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 is crossed out). All axes are in log scale.

work 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 four design goals, and it matches the behavior of several, disparate real social networks. Moreover, we show that NETTIDE is able to forecast the growth of WeChat almost *2 years in the future* (730 days), with $\approx$ 3-percent error, which is impressive given the fact that our model is non-linear and that the "butterfly effect" holds. As points of reference, the weather forecasting (also governed by non-linear equations) goes up to 5-10 days[15], and most forecasting papers in the kdd literature usually do just 1 step look-ahead[25, 29].

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

- **Novel model NETTIDE**: It matches all design goals (G1-G4), that is, it is parsimonious, it includes past models as special cases, it provides the first-ever difference equation for link growth, and it is intuitive.
- **Accuracy:** NETTIDE accurately fits the growth of several, diverse, real networks.
- **Usefulness:** NETTIDE gives excellent forecasting, down to *3%* error, for almost *2 years* ahead in the future.

Reproducibility: Several of the datasets are public[1, 18]; our code is open-sourced at github.com/calvin-zcx/NetTide

The outline of the paper is the typical one: survey, proposed method, experiments, and conclusions.

2. RELATED WORK

We presented related work in two areas: evolving network and growth models.

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 [24], densification growth [24] 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 [17, 21, 24, 23, 9].

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 [34, 4], 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.

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 [27] in the diffusion of innovations. They generate S-shaped sigmoid 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, 32], or the fizzling patterns in the social networks [26]. Models like PhoenixR [13] tried to introduce fizzling mechanism based on Susceptibel Infected Recovered (SIR) model [3]. 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 [22], like Crane-Sornette (CS) model [11], SpikeM [30]. 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.

We summarize the relative advantages and failures of all above

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 $e(t)$			✓						✓

models in Table 1. Only our NETTIDE model meets all the advantages.

3. THE NETTIDE MODEL

Table 2: Symbols and Definitions

Symbols	Definitions
N	Total number of the whole 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 power law sparsity exponent
α	The linear sparsity coefficient

3.1 Preliminaries

Traditional models, like SI/Bass fail to match reality, like the power-law early growth, as shown in Figure 1. We will use the term *early growth* to indicate the time period that is way before the inflection point - in that stage, a sigmoid model exhibits exponential growth. 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 data of Figure 1a and 1b: it can only generate sigmoids over time, which lead to near-exponential initial growth - **not** power-law!

Some reasonable, but wrong attempts: Clearly, we need to replace the SI model with a better one. How should one go about it? 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:

- Unsuccessful attempt 1: *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:

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

  but none of the combinations we tried, gave the the power-law growth of Figure 1.

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:

- Unsuccessful attempt 2: *radioactive decay*

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

No, this does not produce a power-law growth, either. Maybe we should use a few more parameters? After several other attempts, that we will not bore the reader with, the final answer is that we need only *one* additional parameter, provided that we have the correct functional form!

3.2 NETTIDE-Node

We saw that "partial participations" and "radioactive decay" are reasonable, but wrong approaches. It turns out that a good, parsimonious model has: full participations, but *power law* decay of the infectivity/enthusiasm β.

Our G4 goal is intuitiveness - why would human interest fizzle, following a power law? 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 between zero-crossings follows power law with exponent -1.5). Thus, power law decays are as equally justifiable and intuitive as exponential (= radioactive) decays.

And, as we show next, that is all that is needed, to generate the power-law growth of Figures 1a and 1b. Next, we give the details and the proofs for the growth of nodes (NETTIDE-Node model) - 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:

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

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 exponent θ of the power law decay that leads to various growth dynamics, including the power law growth of Figures 1a and 1b. This is the reason that we refer to θ as the *temporal fizzling exponent*.

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

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

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

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. ■

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 equation (7) **Fizzle-Logistic** growth, and equation (8) **Fizzle-Exponential** growth.

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:

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

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 Fizzle-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 Fizzle-Logistic distribution as

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

, 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 $x \ge \frac{t_0^{1-\theta}}{\theta - 1}$,

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

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 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. ■

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 $\frac{dn(t)}{dt} = n(t_0) + \sum_{t_i \le t} \mu_i \frac{1}{(t-t_i)^\theta}$ 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 $\ge$ 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 fizzle-exponential early-growth, and the exponential early-growth as a special case.

3.3 NETTIDE-Link

The growth of social network can never be limited to nodes only. There are no such differential equations to describe the growth dynamics of links before. Here, we give NETTIDE-Link to capture 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:

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

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. For each existing node, he 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 $1+\gamma$.

3.4 NETTIDE parameter learning

The NETTIDE for node and link together has a parsimonious set of parameters, namely, β, θ,β', α, γ and 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: $\sum_{t=t_0}^{T}(n(t)-n^*(t))^2$, by using the *Levenberg-Marquardt* algorithm (LM) [28], which is widely used to solve non-linear least squares problems [30, 13, 33]. 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: $\sum_{t=t_0}^{T}(e(t)-e^*(t))^2$.

3.5 Microscopic explanation

Our NETTIDE can be explained on microscopic level easily. 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 $Edge(t_0)$. For any existing node i in $Node(t)$ at time t, i tries to activate each of his neighbors, like j in G_0.

- *Node growth.* If j has not existed in G_1 yet, then i tries to invite j to join with probability p ($p = \frac{\beta N}{\langle k\rangle t^{\theta}}dt$, where $\langle k\rangle$ is the average degree) in the small time interval $[t, t+dt)$ ($dt = 1$ for the discrete case). If success, we add j to the $Node(t)$ with timestamp 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 q ($q = \frac{\beta'}{t^{\theta}}dt$) in the small time interval $[t, t+dt)$. If success, we add (i, j) 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.

This process is one of the plausible microscopic explanation to describe the growth dynamics of nodes and links. It can be used and extended flexibly, to allow for the structure properties (e.g. with whom to linked in Link growth step) and the user activities (in Activity step). For reproducibility, we open our code, see Section 5.

4. EXPERIMENTS

In this section, we evaluate the effectiveness of NETTIDE on a range of real growing social networks at large scale. Here we report experiments to answer the following questions:

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

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

4.1 Datasets

WeChat on-line social network. WeChat is one of the largest on-line social network in China, which was claimed to have more than 653 million monthly active users by September 30, 2015. 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 the 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.

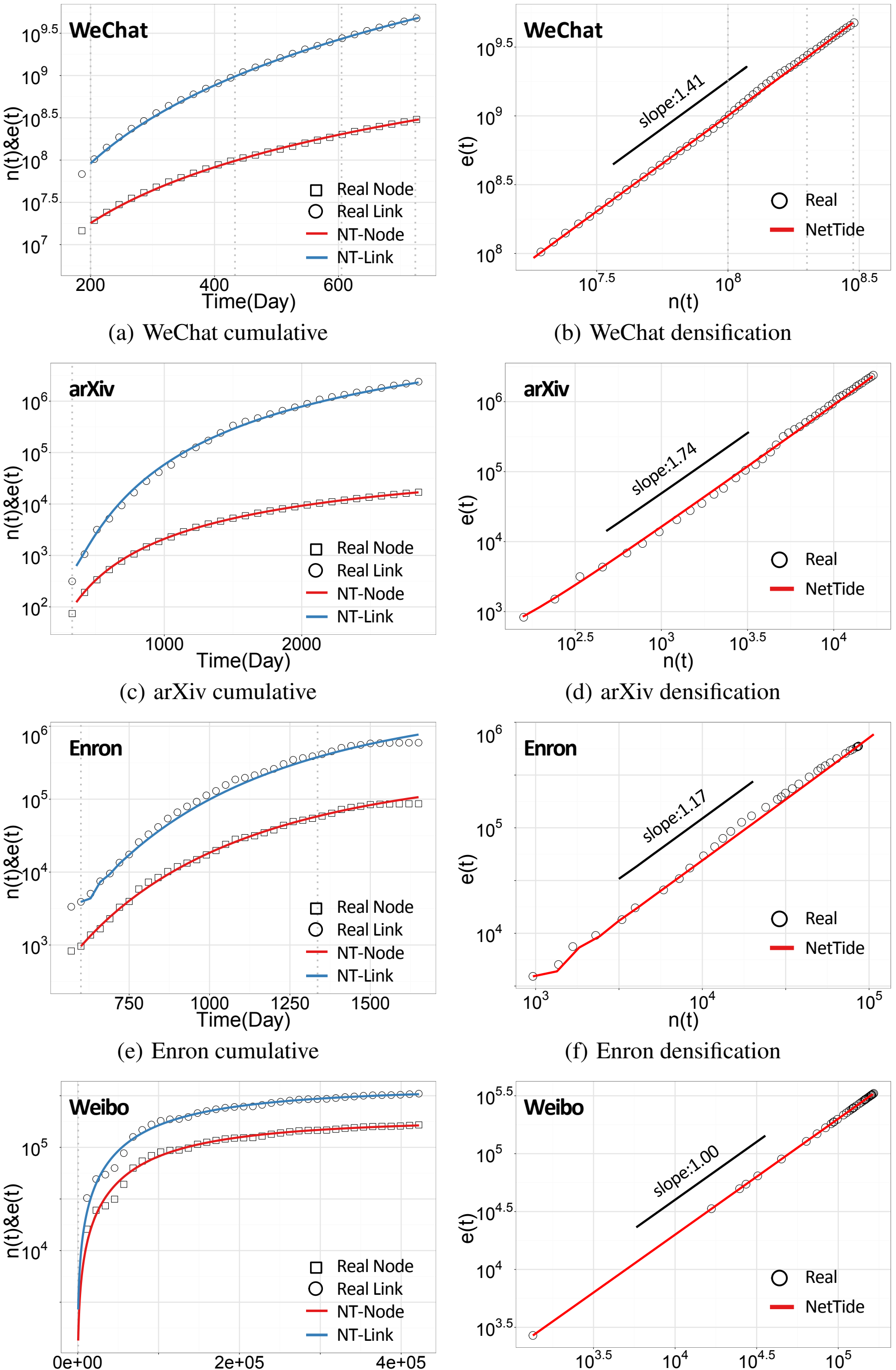
Enron enterprise social network. Through the email records of Enron [18], we recover the enterprise social network emitted from the staff of Enron. The dataset covers the period from January 1998 to July 2002, during which Enron bankrupted on December 2, 2001, causing a sharp cut-off of the $n(t)$. In 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 [35], 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



(a) WeChat cumulative

(b) WeChat densification

(c) arXiv cumulative

(d) arXiv densification

(e) Enron cumulative

(f) Enron densification

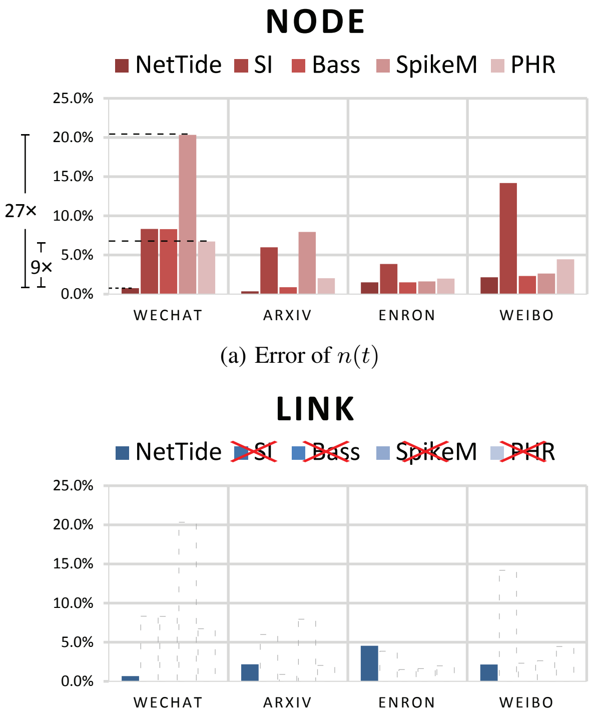
(g) Weibo cumulative

(h) Weibo densification

Figure 2: NETTIDE *fits reality*. Our model fits the growth dynamics of four real social networks very accurately. The four rows corresponds to WeChat (a-b), arXiv (c-d), Enron (e-f) and Weibo (g-h) respectively. In each row, there are three checkpoints: $n(t)$ and $e(t)$ in the first figure, and the $e(n(t))$ in the second figure.

We conduct the experiments in four different real social networks and set three checkpoints to give the empirical evidence for the validity and the generality of our NETTIDE model. The three checkpoints are node cumulative dynamics $n(t)$, link cumulative dynamics $e(t)$, 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 learn the parameters of these baselines the same as our model, i.e. the LM algorithm discussed in Section 3. The microscopic models based on the point process, like Coevolve, are not applicable to our datasets due to their complexity and the need of spreading data. Thus, there is no competitor for link growth dynamics.

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 3 shows the description of the best fitting parameters to four datasets.



(a) Error of $n(t)$

(b) Error of $e(t)$

Figure 3: 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.

Table 3: 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

4.2.2 Value and shape accuracy

Table 4: The comparison of the accuracy of NetTide and baseline methods on three checkpoints of the growth dynamics of four real social networks. With respect to the Normalized-RMSE, NetTide consistently outperforms baselines. All the baselines are not applicable (—) to the links.

WeChat	**n(t)**	**e(t)**	**e(n)**
NetTide	**0.76%**	**0.66%**	**1.08%**
SI	8.32%	—	—
BASS	8.31%	—	—
SPIKEM	20.33%	—	—
PHR	6.73%	—	—
arXiv	**n(t)**	**e(t)**	**e(n)**
NetTide	**0.35%**	**2.18%**	**3.32%**
SI	5.97%	—	—
BASS	0.88%	—	—
SPIKEM	7.95%	—	—
PHR	2.03%	—	—
Enron	**n(t)**	**e(t)**	**e(n)**
NetTide	**1.51%**	**4.54%**	**4.62%**
SI	3.84%	—	—
BASS	1.51%	—	—
SPIKEM	1.63%	—	—
PHR	1.99%	—	—
Weibo	**n(t)**	**e(t)**	**e(n)**
NetTide	**2.15%**	**2.15%**	**0.06%**
SI	14.19%	—	—
BASS	2.31%	—	—
SPIKEM	2.62%	—	—
PHR	4.45%	—	—

Our NetTide accurately fits growth dynamics of both nodes and links of WeChat, which span 726 days from the release. The fitting results of the five checkpoints, as depicted in Fig 2a-b, show that the results of our NetTide almost overlap all the real data points. The fitting covers the period (just after the release of WeChat V2.0, which includes voice notes) during which WeChat gained its major population, and we treat the first 199 days as the burn-in period because of the unstable version update of the WeChat, and a relatively small number of users and links (5.8% of the total nodes and 1.8% of the total links in our data). 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 4). 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 3b).

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 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, ignoring the first 1-percent population as in the burning period. 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 2c-d) and Enron (Fig 2e-f). Specifically, NetTide-Node gets the lowest error 0.35% (1.51%) in the arXiv (Enron) case compared with baselines, as shown in Fig 3a. 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 3b.

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 (as shown in Fig 2g) introduce a relatively large error of $n(t)$ (Table 4), 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 3b.

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 3a, 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 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.

4.3 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 4a. In arXiv case, they are the dates of reaching 3000, 6000, 9000 authors respectively, in Fig 4c. The same task goes with NETTIDE-Link, in which case the number of links is seldom predicted before.

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 4a). For the arXiv, the overall errors are 2.86% and 4.18% for $n(t)$ and $e(t)$ respectively (Fig 4c) from t_1 to t_3 . As a reference, we compare our forecasting results with 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 4).

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 4a, 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.

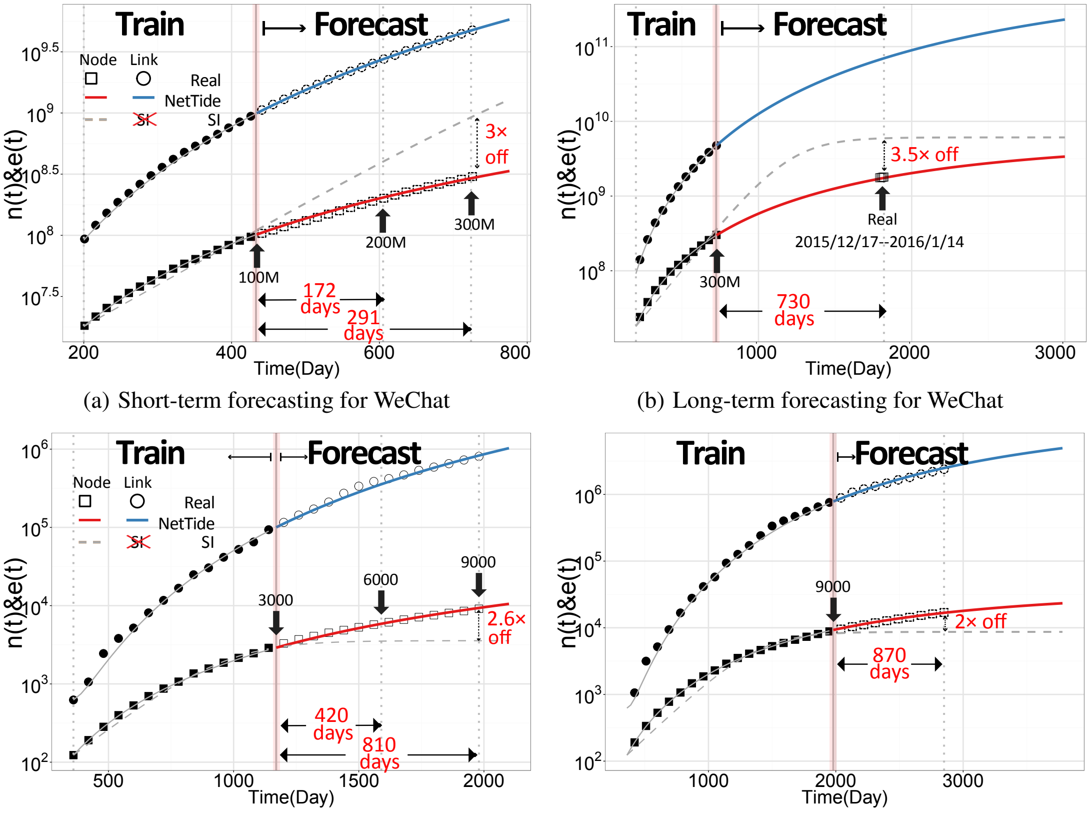
In all, our NETTIDE achieves a surprisingly high forecasting accuracy for both node and link growth in the short term.

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 4b). 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 to 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

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



(a) Short-term forecasting for WeChat

(b) Long-term forecasting for WeChat

(c) Short-term forecasting for arXiv

(d) Long-term forecasting for arXiv

Figure 4: 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, and the below is arXiv. (a) and (c) are the results of the short term forecasting, while (b) and (d) are the results of the long term case.

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 4d. 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.

5. CONCLUSIONS

In this paper, we studied the growth dynamics of real social networks and presented NetTide to capture both node and link growth dynamics. We examine a range of real evolving social networks, especially China's largest on-line social network WeChat, and find that both node and link in real social networks follow near power law growth dynamics, rather than the exponential early 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 the unified but parsimonious model to capture real social network growth, like the power law early-growth of the Log-Logistic, and the general form Fizzle-Exponential early-growth of the Fizzle-Logistic. Our NetTide-Link is the first-ever differential equation to capture the growth dynamics of links, accurately fitting reality. The main contributions are:

1. **Novel model NetTide**: NetTide-Node captures a large range of real growth dynamics and NetTide-Link is the first differential equation to capture the link growth dynamics. Both equations are parsimonious and explainable on micro level.
2. **Accuracy**: We presented experiments on four real evolving social networks, especially the WeChat (300 million nodes, 4.75 billion links). Our NetTide model matches the real growth patterns accurately.
3. **Usefulness**: Our NetTide can be used to do 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).

Reproducibility: We have already open-sourced our code of the NetTide process to generate the growth dynamics of both nodes and links, at https://github.com/calvin-zcx/NetTide. Several of the datasets are public[1, 18].

Acknowledgments

This work was supported by National Program on Key Basic Research Project, No. 2015CB352300; National Natural Science Foundation of China, No. 61370022, No. 61531006, No. 61472444 and

No. 61210008. Thanks for the research fund of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology. This material is based upon work supported by the National Science Foundation under Grant No. CNS-1314632 IIS-1408924 Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, or other funding parties.

6. 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] F. Hamilton, T. Berry, and T. Sauer. Predicting chaotic time series with a partial model. *Physical Review E*, 92(1):010902, 2015.

[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] 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.

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

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

[21] 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.

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

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] Y. Matsubara, Y. Sakurai, and C. Faloutsos. Autoplait: Automatic mining of co-evolving time sequences. In *SIGMOD'14*, pages 193–204. ACM, 2014.

[30] 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.

[31] 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.

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

[33] 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.

[34] 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 *KDD'13*, pages 356–364. ACM, 2013.

[35] 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 ICDM*, 2015.