Uncovering and Predicting the dynamic process of Information Cascades with survival model

Linyun Yu[1] Peng Cui[1] Fei Wang[2] Chaoming Song[3] Shiqiang Yang[1]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China

[2]Department of Computer Science and Engineering, University of Connecticut, USA

[3]Department of Physics, University of Miami

Information Cascade

Predicting these cascades are very important for many applications.

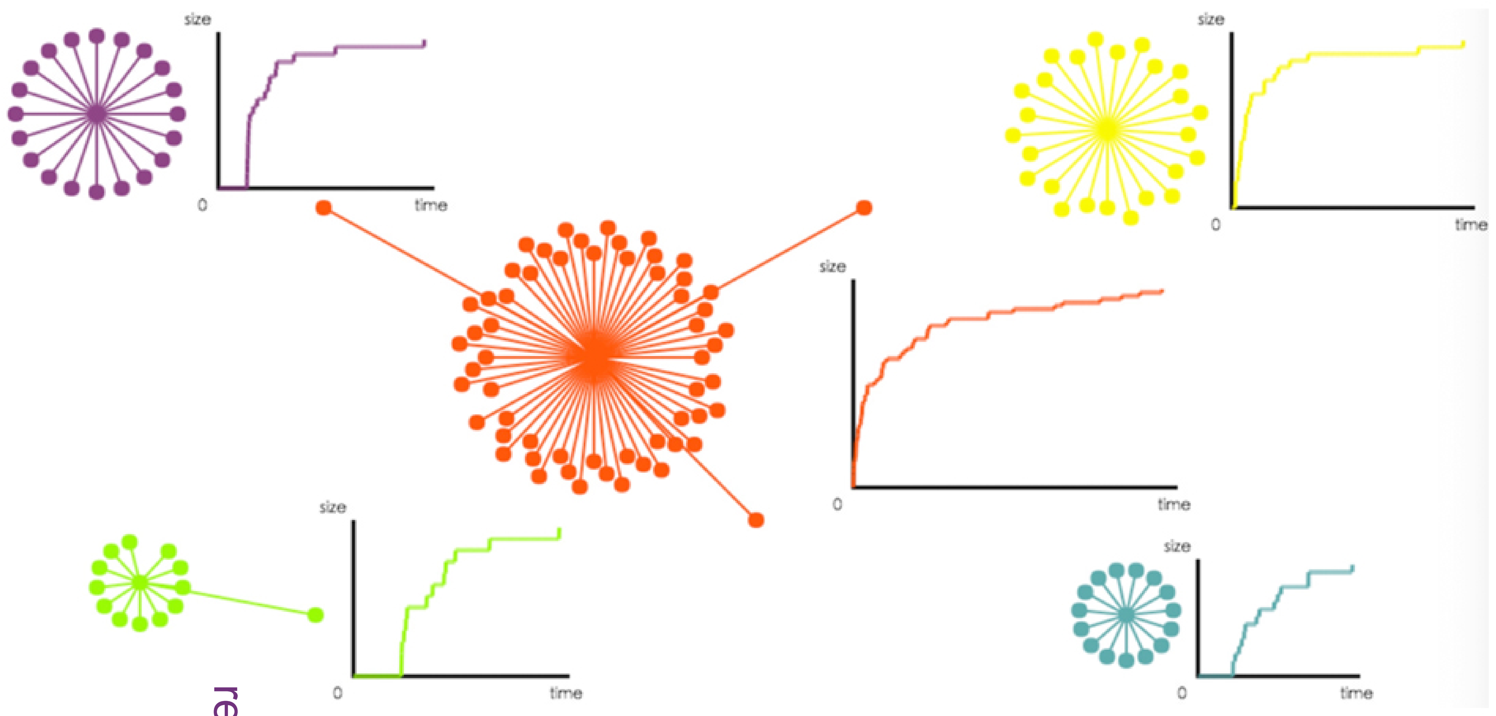
It is changing all the time

In network environment, if decentralized nodes act on the basis of how their neighbors act at earlier time, cascades will be formed.

decomposition

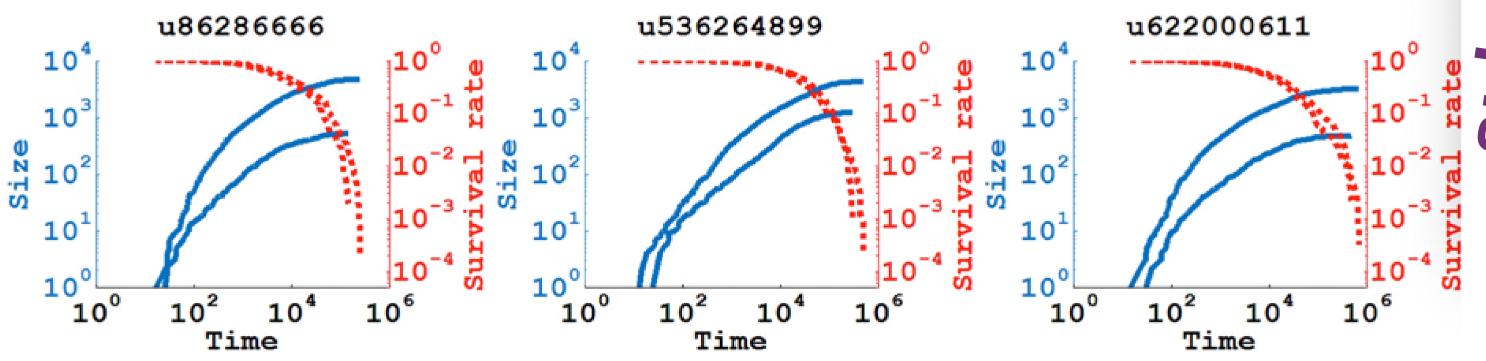
combination

Subcascades



representation

User behavioral dynamics



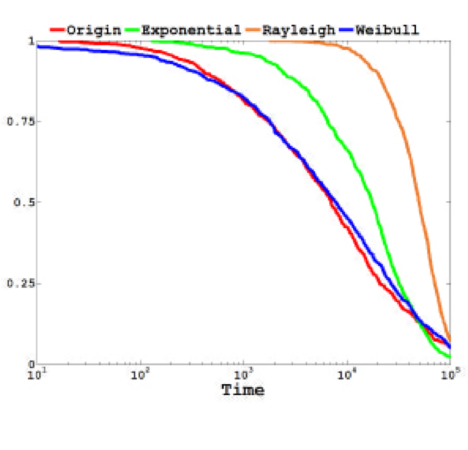
·Averaging the size growth curve:

Different subcascades of the same user might have different size growth curves.

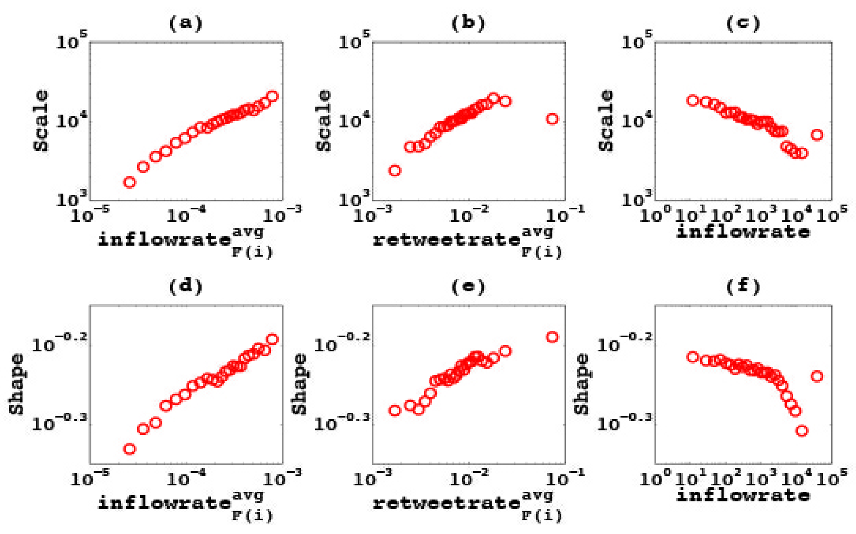
·Survival rate: the percentage of nodes that has not been but will be infected

For different subcascades of the same user, the survival function is quite stable.

Parameterized Behavioral Dynamics



Covariates of Behavioral Dynamics



NEtworked WEibull Regression (NEWER)

$$F(\lambda, k, \beta, \gamma) = G_1(\lambda, k) + \mu G_2(\beta, \lambda) + \eta G_3(\gamma, k)$$

$$G_1(\lambda, k) = -\log L(\lambda, k)$$ Log likelihood

$$G_2(\lambda, \beta) = \frac{1}{2N} \|\log \lambda - \log X \cdot \beta\|^2 + \alpha_\beta \|\beta\|_1$$ Parameterize λ

$$G_3(k, \gamma) = \frac{1}{2N} \|\log k - \log X \cdot \gamma\|^2 + \alpha_\gamma \|\gamma\|_1$$ Parameterize k

- ☐ Theoretically proved to be lower-bounded.
- ☐ Coordinate Descent strategy is exploited with guaranteed convergence.

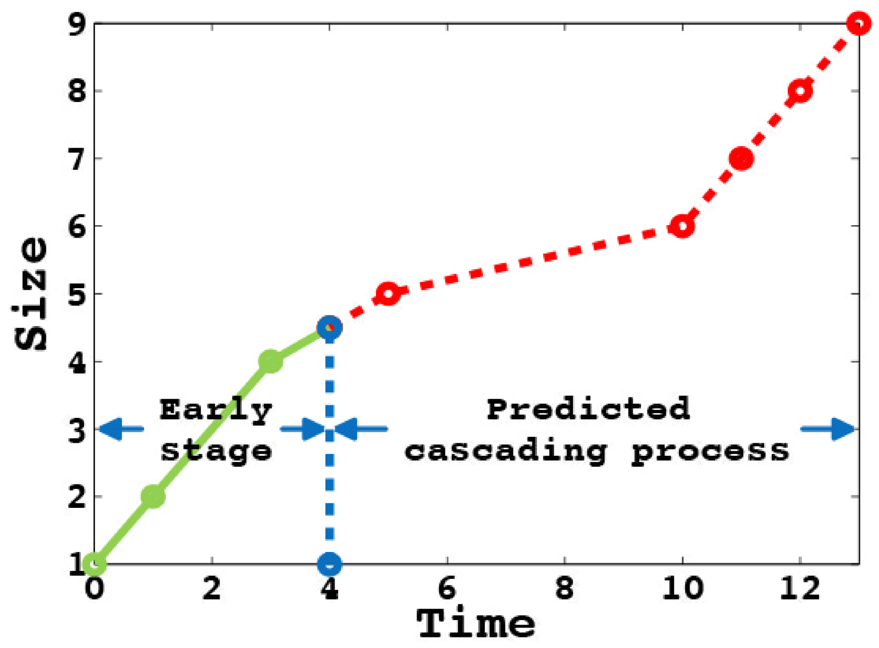
Contact

Name: **Linyun Yu 余林韵**

Mail: mail@scalerhyme.com

https://www.scalerhyme.com

Cascading Process Prediction



Source: the early stage of an information cascade.

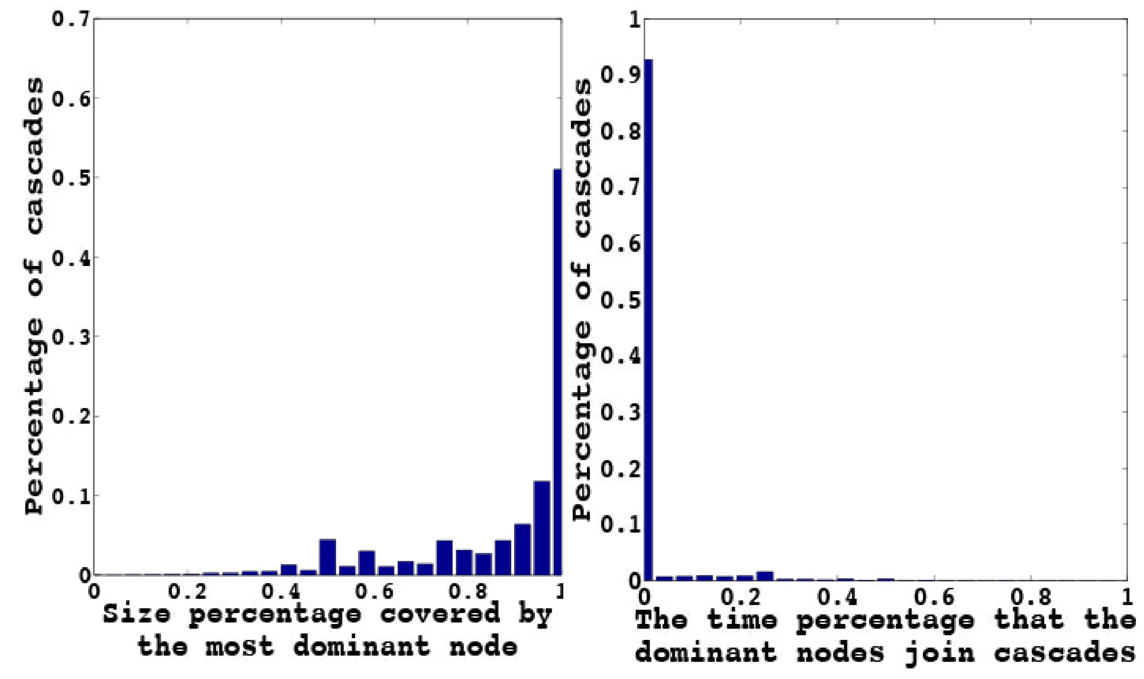
Target: the later stage of the information cascade, or its cumulative cascade size of any later time.

Make predictions

$$CascadingProcess = \sum_{\text{all observed subcascades}} SubcascadeProcess(S)$$

subcascade prediction

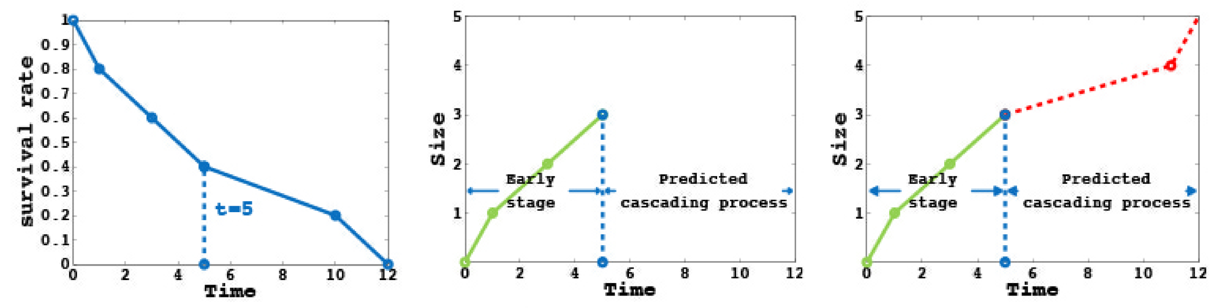
Observations supporting ignoring unobserved subcascades



·A very small number of nodes whose behavioral dynamics dominate the cascading process.

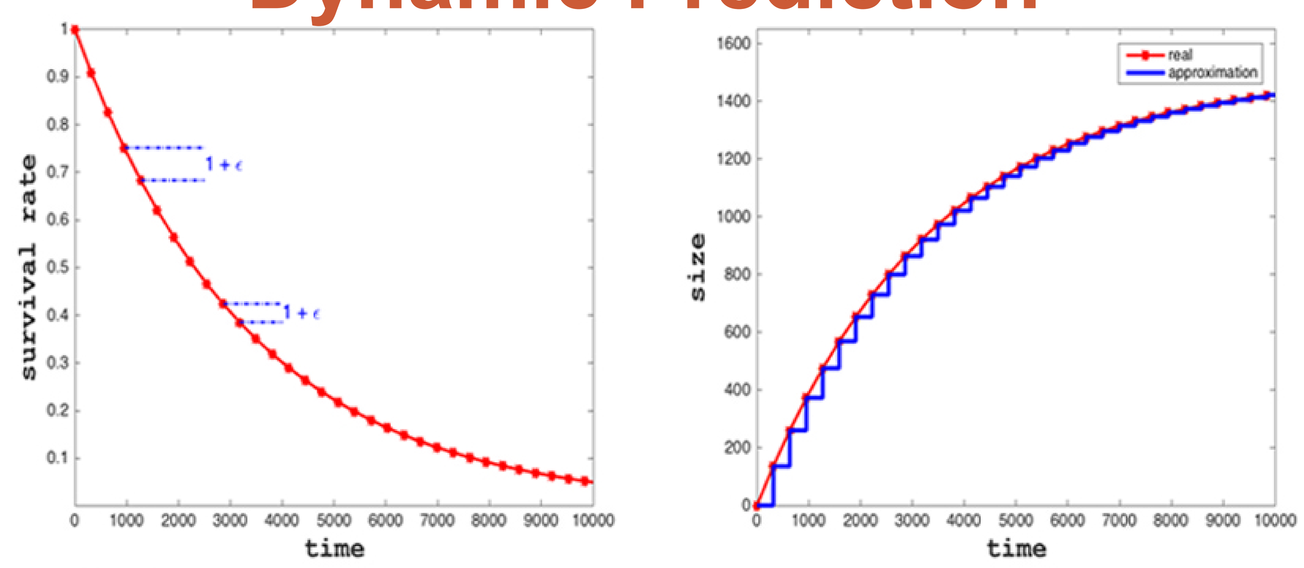
·Most of dominate nodes in actual join cascades in the very early stage.

From rate dimension to size dimension



Rate - - - - Early Stage Information - - - -> Size

Dynamic Prediction



·Dynamic prediction:

·ignore most recalculations for subcascades by using the previous calculation instead.

·Setting the calculation time point based on the last calculation.

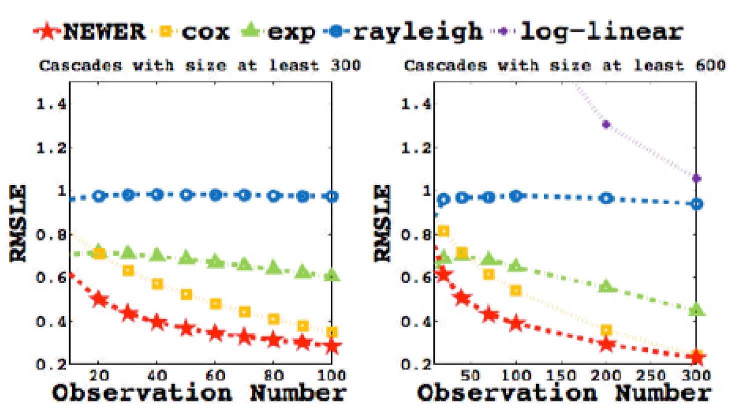
·Subcascade process prediction:

·carve up the subcascade process based on the survival function

·maintain subcascade process using balanced binary search tree (BBST)

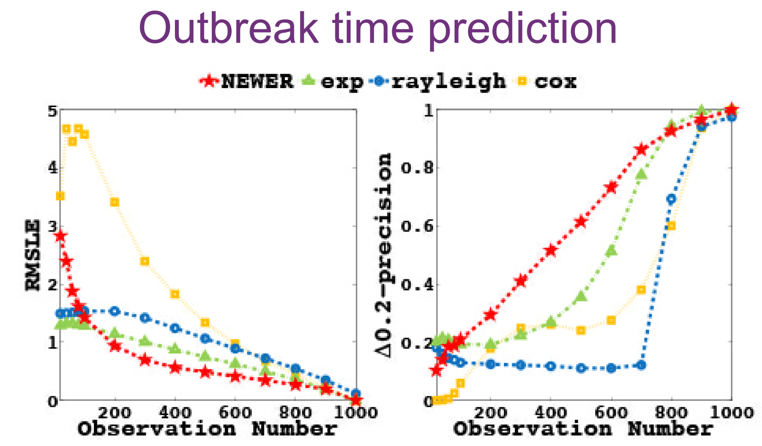
Evaluation Result

Cascade Size Prediction



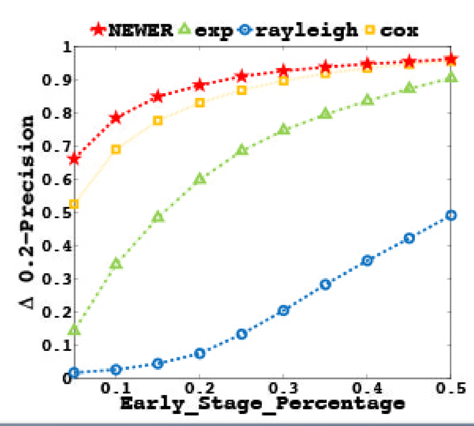
The improvements are more obvious in early stage.

Outbreak time prediction



Average improvements in early stage: 70%

Cascading Process Prediction



90% percentage accuracy when we have only 20% early stage informations.

Efficiency of the Method

Method	Without Sampling Strategy	With Sampling Strategy ($\delta = 0.1$)
Size ≥ 20	$8.47 * 10^5 s$	$10.73s$
Size ≥ 50	$7.61 * 10^5 s$	$8.62s$
Size ≥ 100	$6.65 * 10^5 s$	$7.09s$
Size ≥ 500	$4.35 * 10^5 s$	$4.33s$
Size ≥ 1000	$3.4 * 10^5 s$	$3.30s$

Running time for cascade size prediction

improvements: 10^5

Size	Without Sampling Strategy	With Sampling Strategy ($\epsilon_1 = 0.1$ and $\epsilon_2 = 0.1$)
20	$1.4 * 10^{12}$	$4.2 * 10^6$
50	$3.5 * 10^{12}$	$7.6 * 10^6$
100	$6.9 * 10^{12}$	$1.4 * 10^7$
500	$3.5 * 10^{13}$	$3.4 * 10^7$
around 1000	$6.9 * 10^{13}$	$4.2 * 10^7$

Calculation number for cascade process prediction

improvements: 10^6