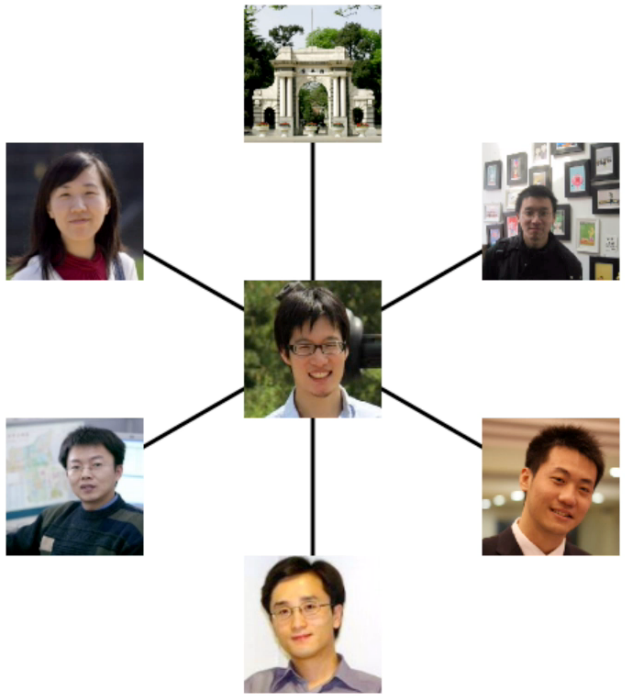
From Micro to Macro: Uncovering and Predicting Information Cascading Process with Behavior Dynamics

Linyun Yu, Tsinghua University, Beijing, China

Peng Cui (Tsinghua), Fei Wang(UNIVERSITY OF CONNECTICUT), Chaoming Song (UNIVERSITY OF MIAMI), Shiqiang Yang (Tsinghua)

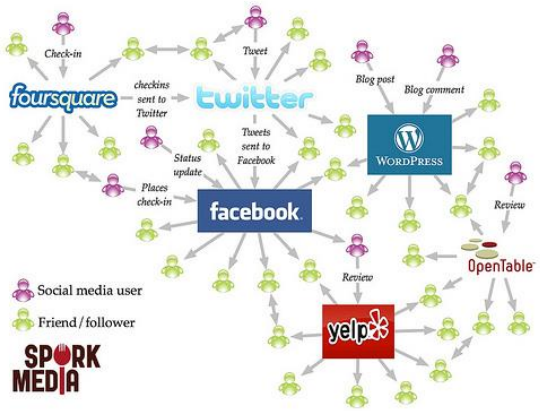
Information Cascade

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



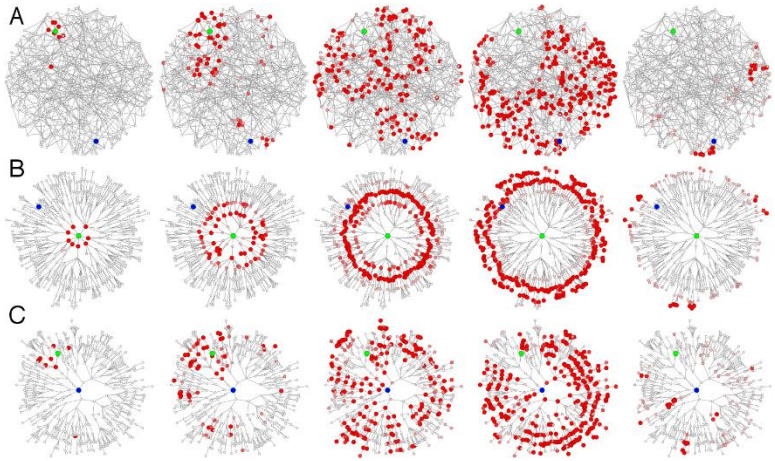
Information spreading is ubiquitous

Social Media

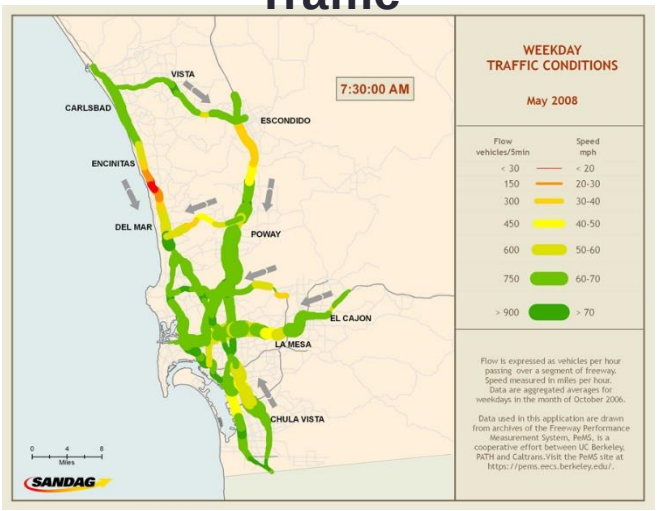


Word-of-Mouth (Marketing)

Epidemics

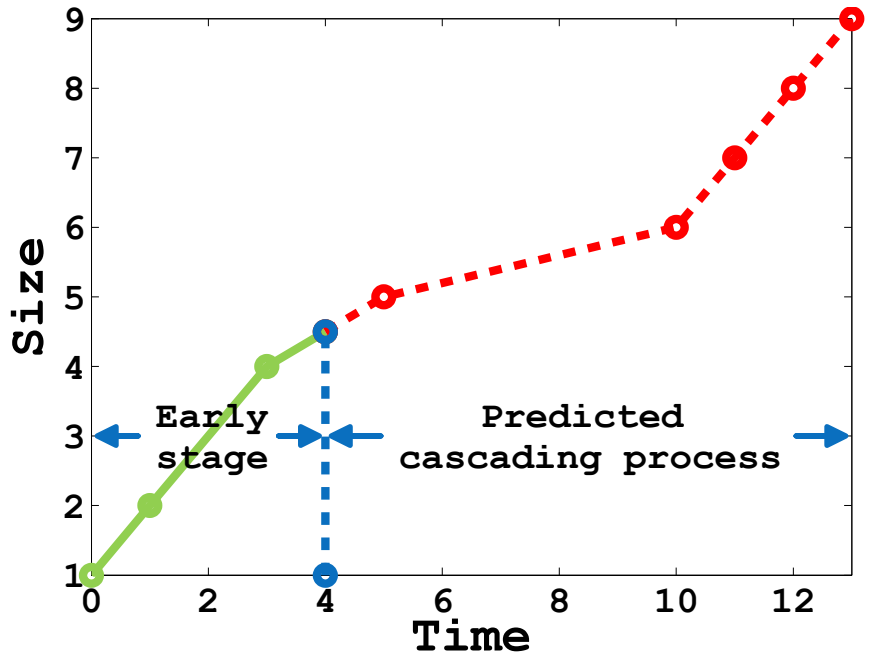


Traffic



Cascading Process Prediction

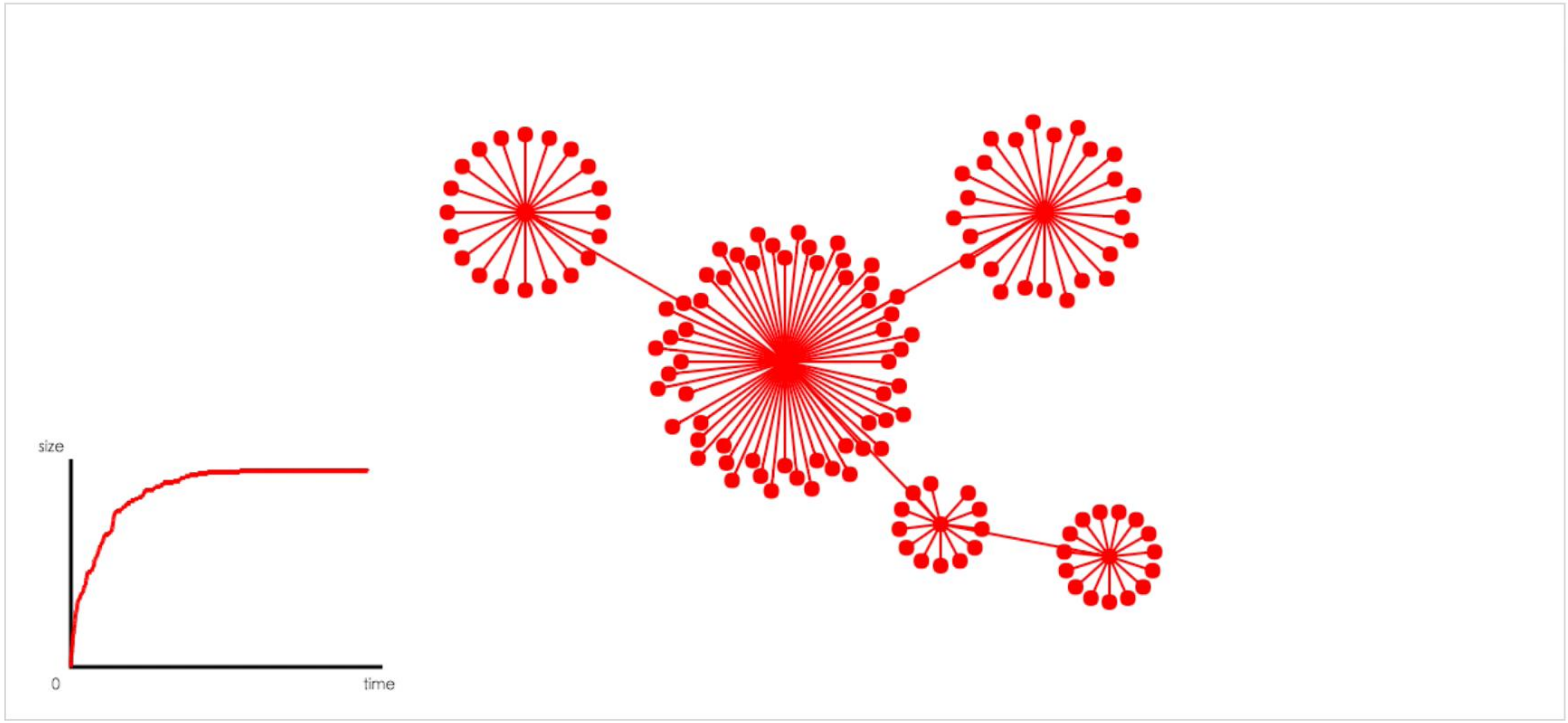
- Problem Definition:
  - Source: the early stage of an information cascade
  - Target: the later stage of the information cascade, or its cumulative cascade size of any later time



Cascading Process

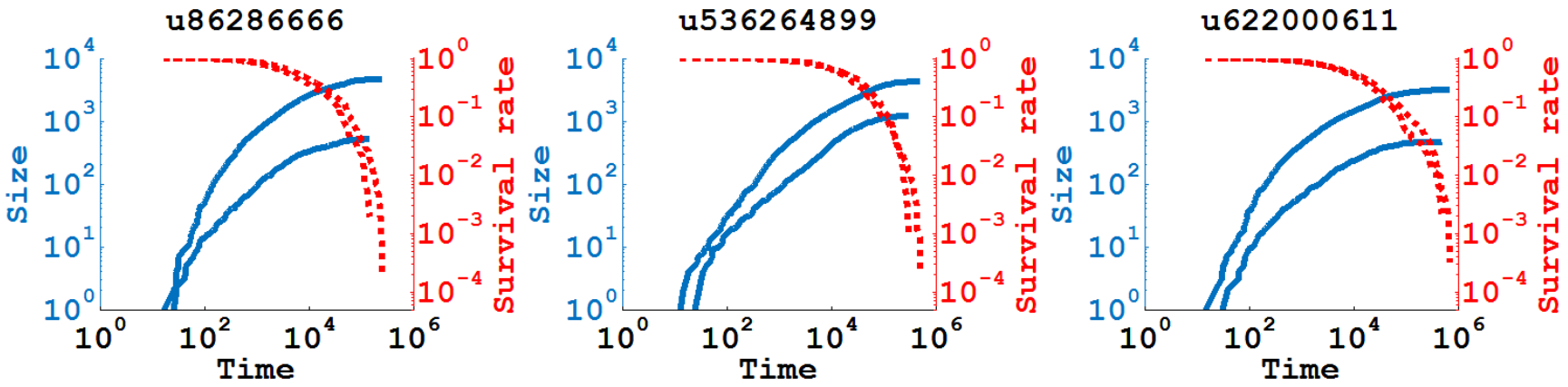
From Macro to Micro: Subcascades

- How to model subcascades?
- How to connect subcascades and the global cascade?
- How to make predictions fast and accurate?



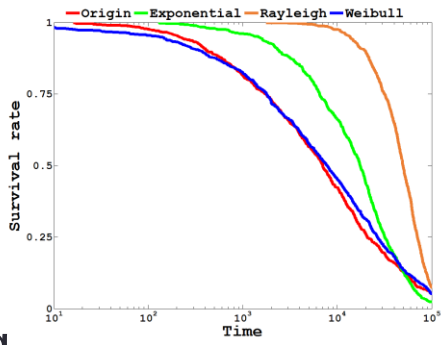
User Behavioral Dynamics

- **Behavioral Dynamics of a user**: The changing process of its offspring nodes that involve in the cascade after the user involved in the post.
- **Representation**
  - 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.



Parameterize Behavioral Dynamics

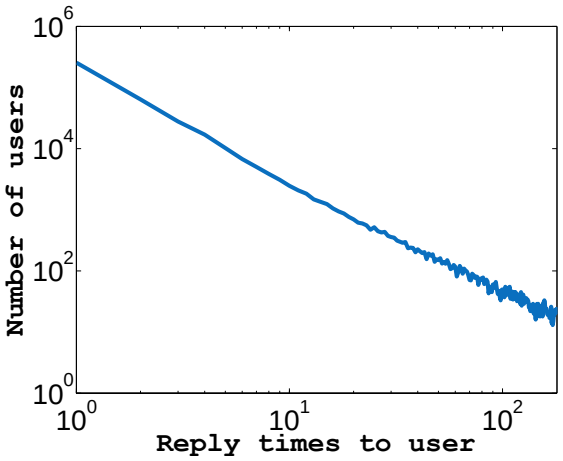
- The behavioral dynamics need to be parametrized for the ease of computation and modeling.
- Exponential and Rayleigh distributions cannot well capture both the scale and shape characteristics of behavioral dynamics.
- The Weibull distribution is adequate for parameterizing behavioral dynamics:
  - λ: control the scale parameter
  - k: control the shape parameter

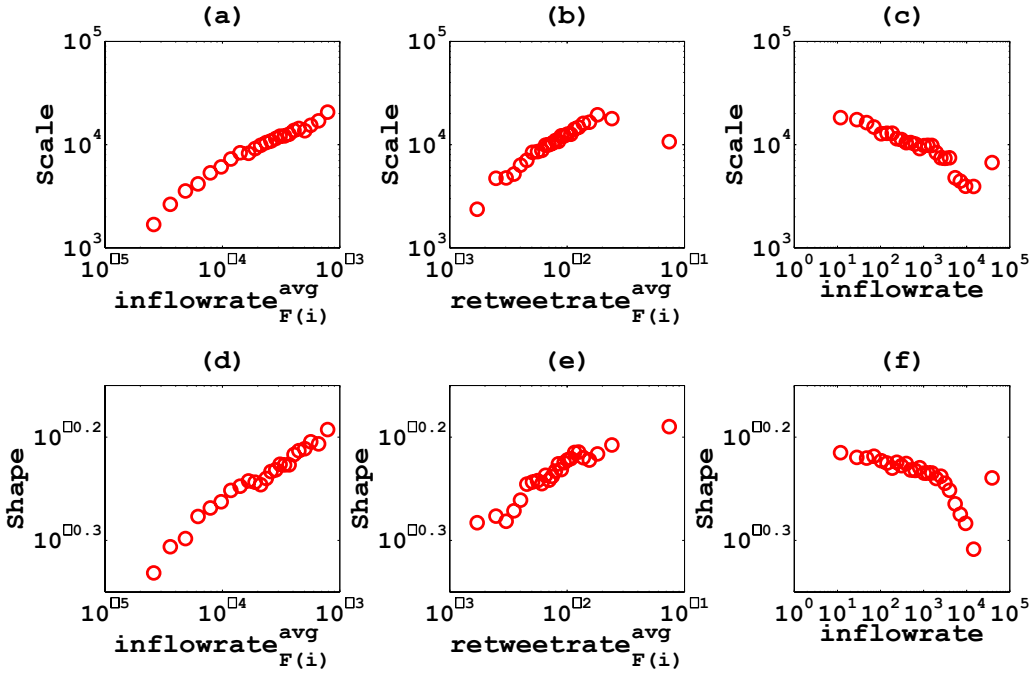


model	density function	survival function	hazard function	ks-static in Weibo
Exponential	$\lambda_i e^{-\lambda_i t}$	$e^{-\lambda_i t}$	λ_i	0.2741
Power Law	$\frac{\alpha_i}{\delta}\left(\frac{t}{\delta}\right)^{-\alpha_i-1}$	$\left(\frac{t}{\delta}\right)^{-\alpha_i}$	$\frac{\alpha_i}{t}$	0.9893
Rayleigh	$\alpha_i t e^{-\alpha_i \frac{t^2}{2}}$	$e^{-\alpha_i \frac{t^2}{2}}$	$\alpha_i t$	0.7842
Weibull	$\frac{k_i}{\lambda_i}\left(\frac{t}{\lambda_i}\right)^{k_i-1} e^{-\left(\frac{t}{\lambda_i}\right)^{k_i}}$	$e^{-\left(\frac{t}{\lambda_i}\right)^{k_i}}$	$\frac{k_i}{\lambda_i}\left(\frac{t}{\lambda_i}\right)^{k_i-1}$	**0.0738**

Covariates of Behavioral Dynamics

- Interaction information between nodes is not always available. It is difficult to measure out-of-sample nodes.





The parameters of the user's behavioral dynamics can be well estimated by the behavioral features of its network neighbors.

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

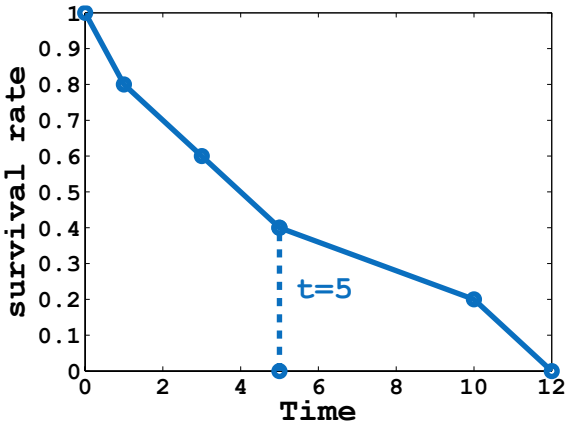
$$G_2(\lambda, \beta) = \frac{1}{2N} \left\| \log \lambda - \log X \cdot \beta \right\|^2 + \alpha_\beta \left\| \beta \right\|_1$$ Parameterize λ

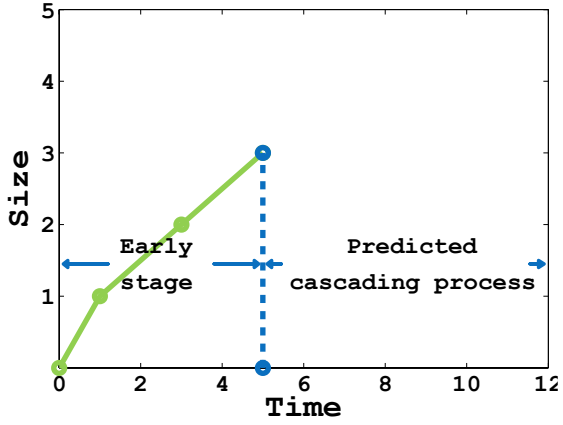
$$G_3(k, \gamma) = \frac{1}{2N} \left\| \log k - \log X \cdot \gamma \right\|^2 + \alpha_\gamma \left\| \gamma \right\|_1$$ Parameterize k

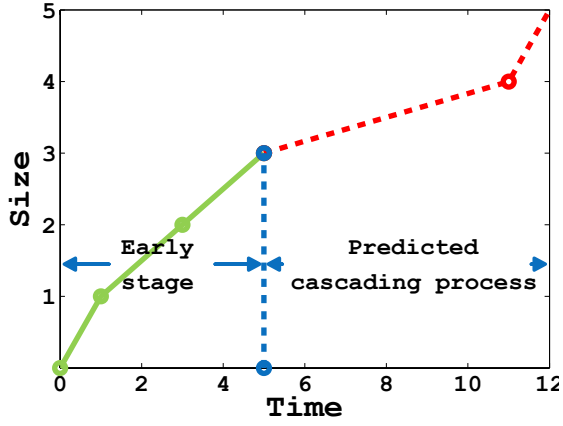
- Theoretically proved to be lower-bounded.
- Coordinate Descent strategy is exploited with guaranteed convergence.

Subcascade Process Prediction

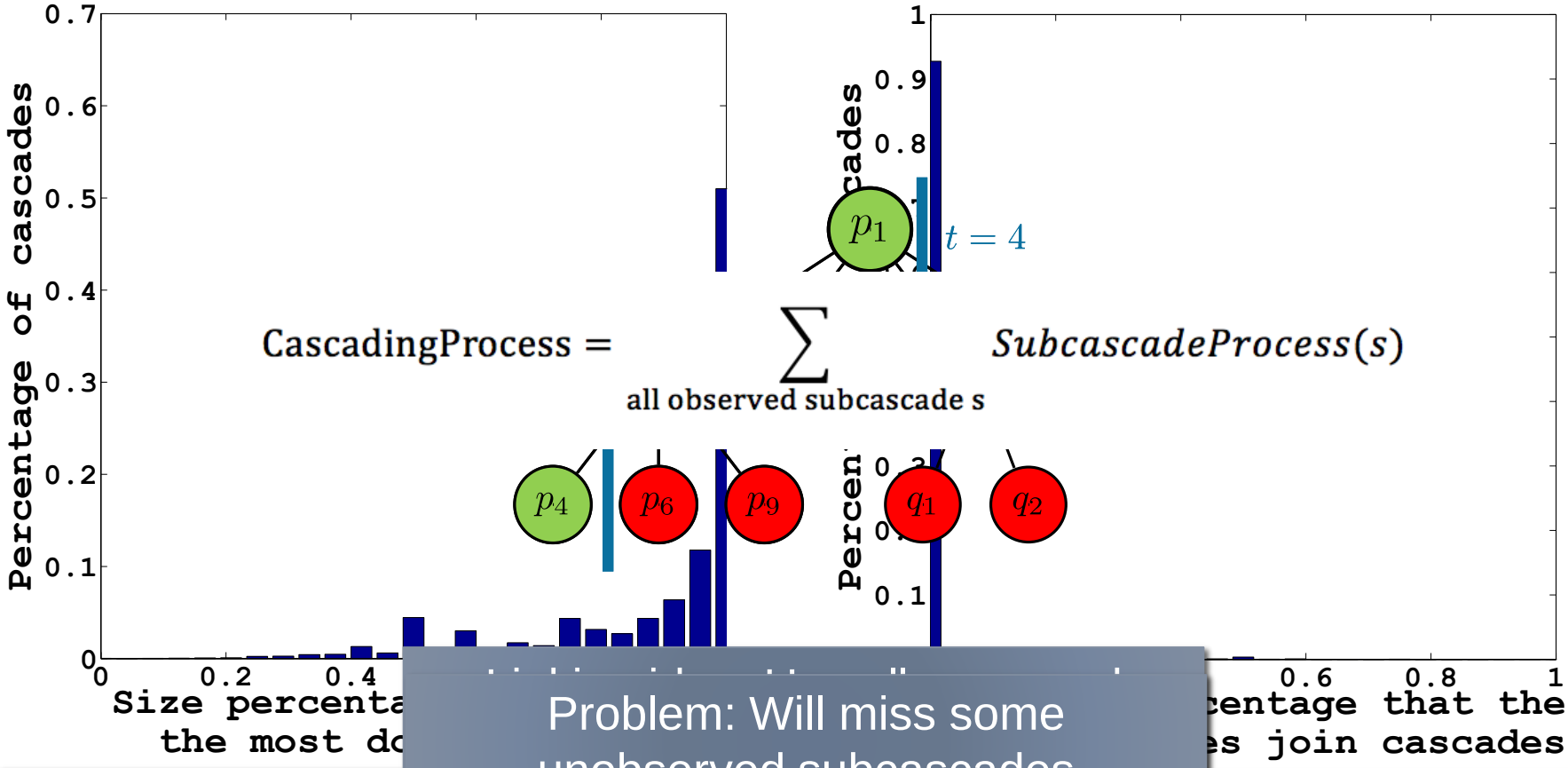
- From rate dimension to size dimension







From Subcascades to Cascade



$$\text{CascadingProcess} = \sum_{\text{all observed subcascade s}} \textit{SubcascadeProcess}(s)$$

Problem: Will miss some unobserved subcascades

Minor dominan… dominate the cascading process.

…inance: Dominant nodes are prone to join early.

Dynamic Prediction

Real Application Demand:

- Accuracy
- Real-time

Sampling strategy:

- Ignore most recalculations for subcascades by using the previous calculation instead.
- Setting the calculation time point based on the last calculation.

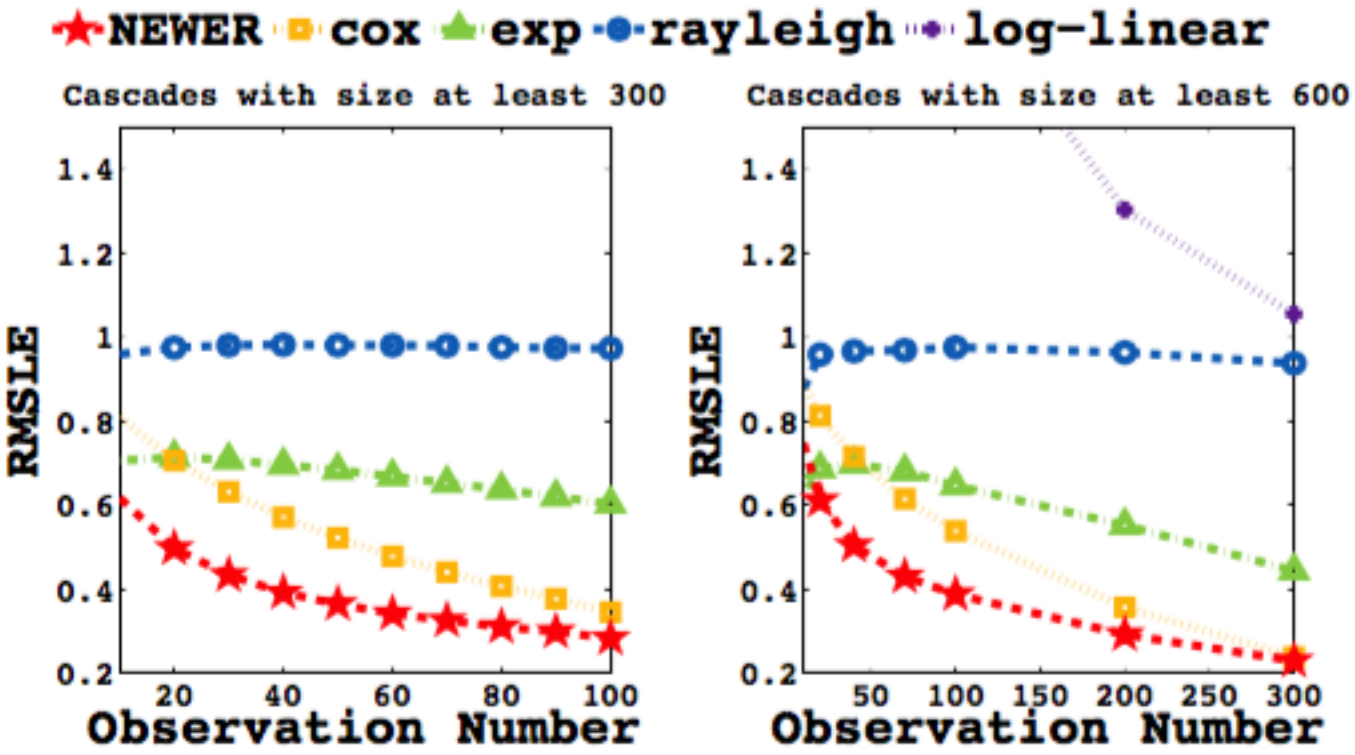
Time complexity: from $O(T^2)$ to $O(T)$ (with an error bound)

Experiments

- Datasets: Tencent Weibo
  - All cascades generated between Nov 15th and Nov 25th in 2011.
  - retain all 0.59 million cascades that the cascades size are at least 5.
- Baseline:
  - Cox Proportional Hazard Regression Model (Cox)
  - Exponential/Rayleigh Proportional Hazard Regression Model (Exponential/Rayleigh)
  - log-Linear regression(Log-linear)
- Evaluation metric:
  - RMSLE: Root Mean Square Log Error
  - $\Delta\sigma$-Precision: Precision value that the predicted value within $(1+\sigma) \pm 1$groundtruth

Cascade Size Prediction

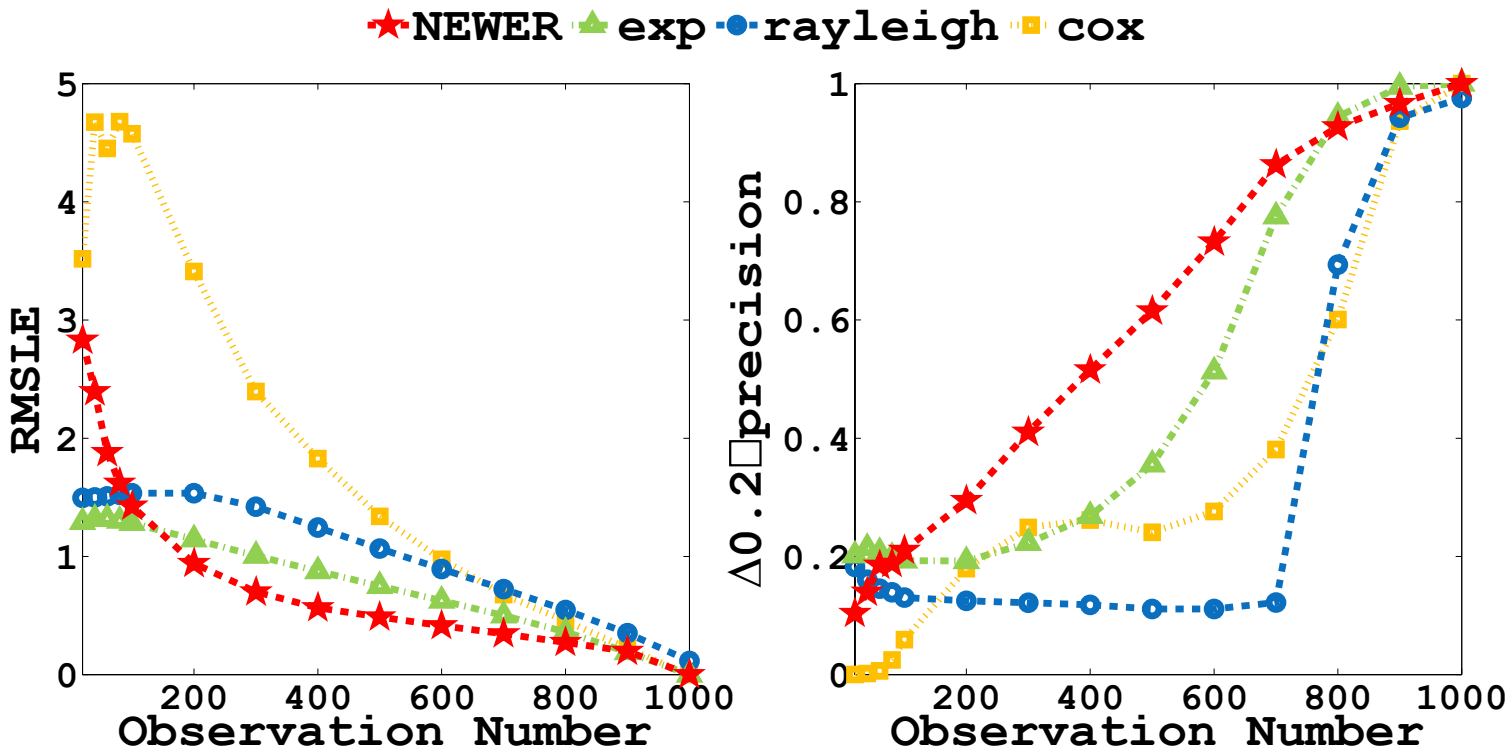
- What is the final size of the cascade?



The improvements are more obvious in early stage.

Outbreak Time Prediction

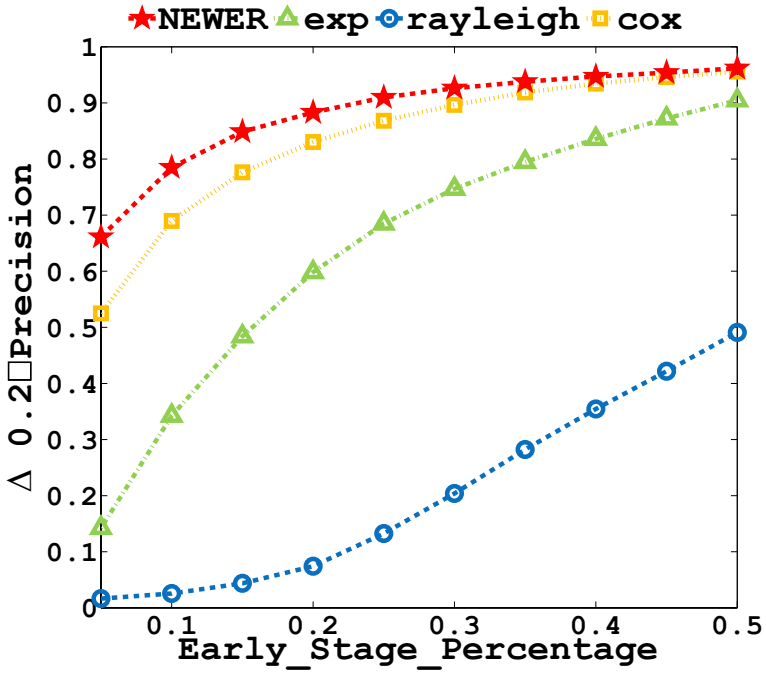
- When will the cascade break out?



Average improvements in early stage:
70%

Cascading Process Prediction

- What is the size of the cascade at any later point?



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

Efficiency of the method

- How fast can our method achieve?

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

Conclusion

- A new Problem:
  - Given early stage information, predict the future cascading process.
- A new angle:
  - Uncover the cascading process through behavioral dynamics.
- A new model (NEWER):
  - Model the behavioral dynamics and predict the subcascading process
- A scalable solution:
  - Predict the dynamic process of information cascade with linear complexity

The proposed method has been transferred to Tencent for social marketing.

Thanks!