Modelling Cascades Over Time in Microblogs

Wei Xie, Feida Zhu, Siyuan Liu and Ke Wang*
Living Analytics Research Centre
Singapore Management University

* Ke Wang is from Simon Fraser University, and this work was done when the author was visiting Living Analytics Research Centre in Singapore Management University.

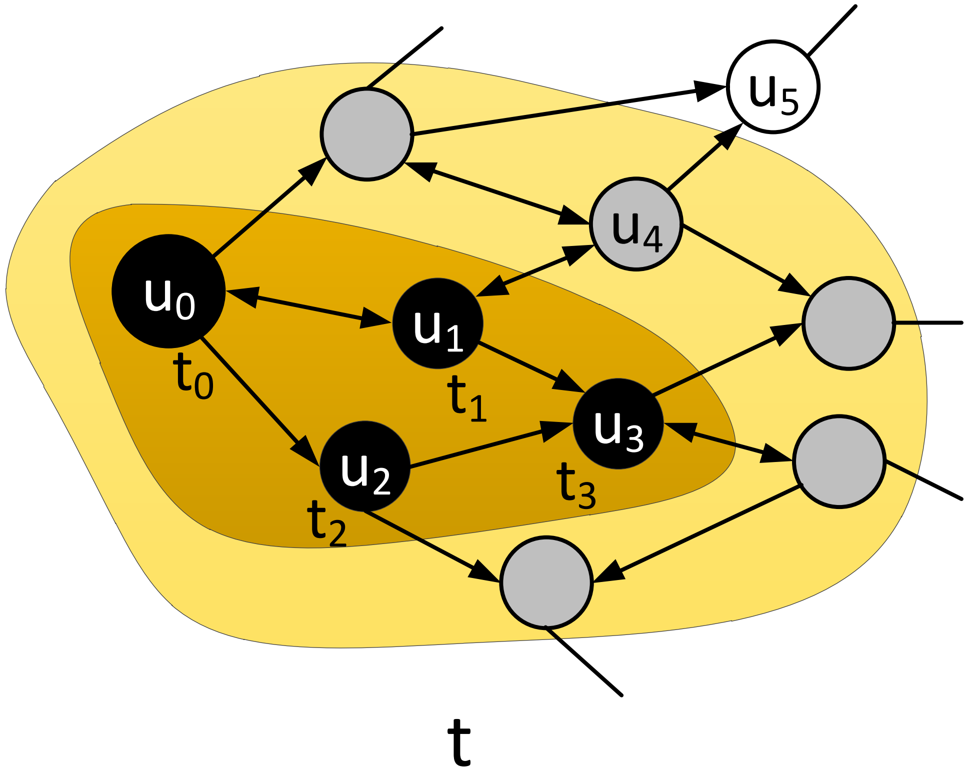
Motivation

- Business applications such as viral marketing have driven a lot of research effort predicting whether a cascade will go viral.
- In real life, there are very few truly viral cascades.
- Previous research work* shows that temporal features are the key predictor of cascade size.

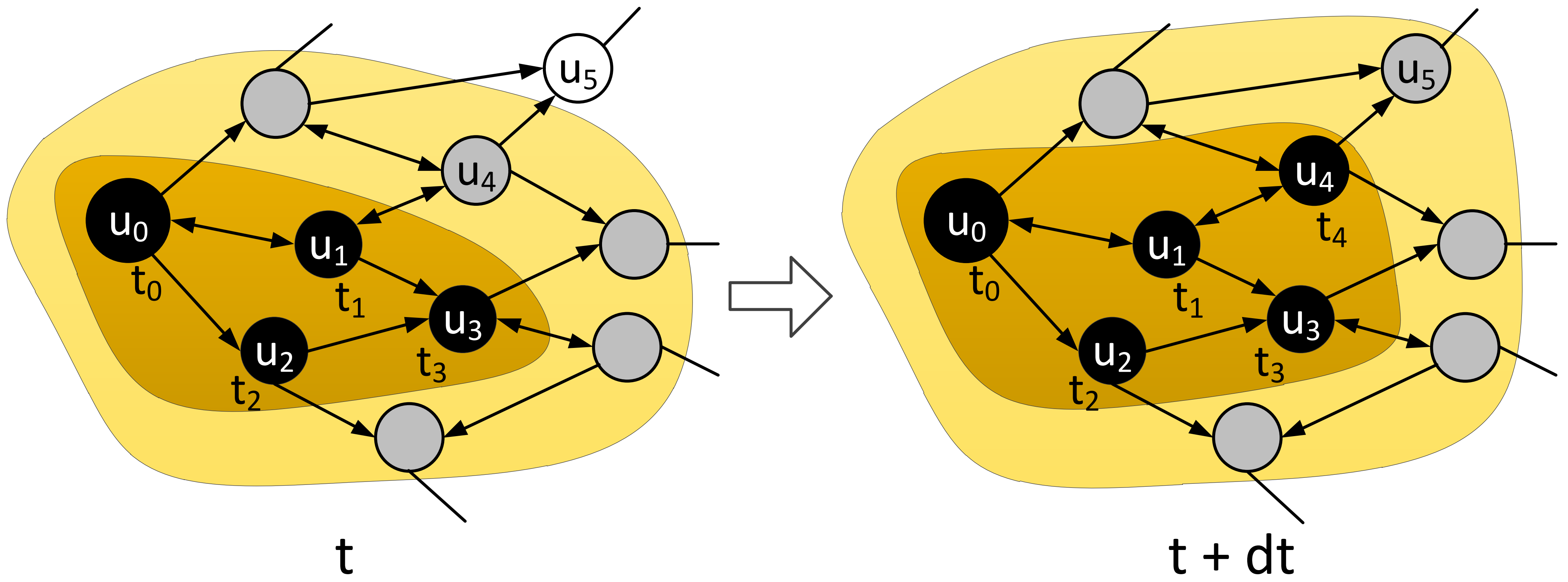
* Justin Cheng, Lada A. Adamic, P. Alex Dow, Jon M. Kleinberg, Jure Leskovec: Can cascades be predicted? WWW 2014: 925-936

LARC
LIVING ANALYTICS RESEARCH CENTRE

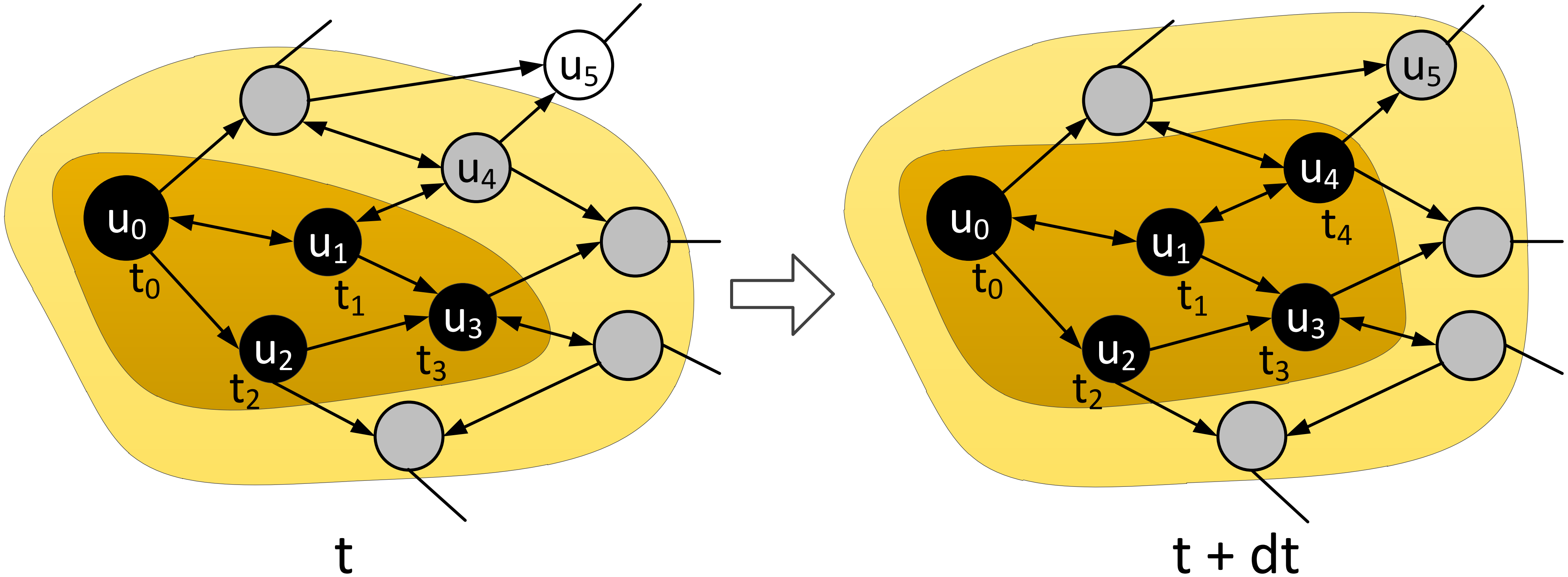
Time-aware Cascade Model



Time-aware Cascade Model



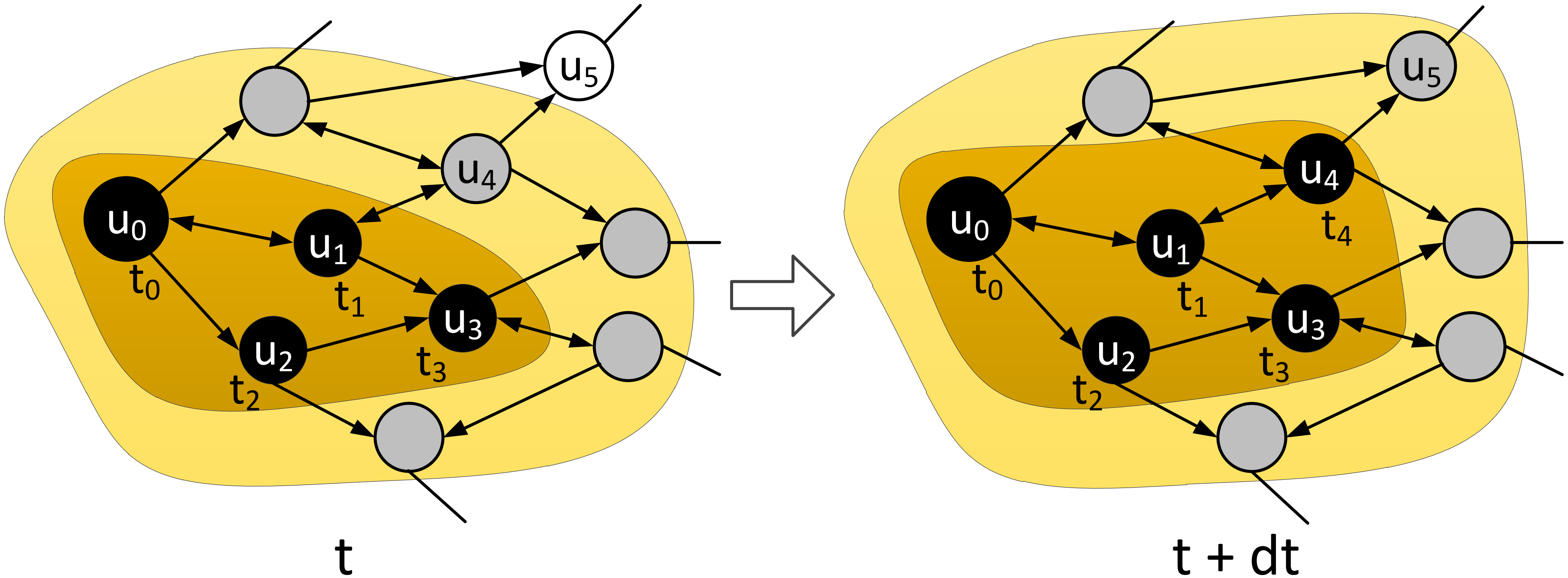
Time-aware Cascade Model



$$P_i(t) = h_i(t, \{t_j\}_{u_j \in Followee^{(i)}(t)}; \mathbf{\Theta}) \cdot dt$$

$$\begin{cases} P(\mathbb{C}(t+dt)) = P(\mathbb{C}(t+dt)|\mathbb{C}(t)) \cdot P(\mathbb{C}(t)) \\ P(\mathbb{C}(t_0)) = 1 \\ P(\mathbb{C}(t+dt)|\mathbb{C}(t)) = \prod_{u_i \in \mathbb{X}^{(1)}(t)} P_i(t) \cdot \prod_{u_{i'} \in \mathbb{X}^{(2)}(t)} (1 - P_{i'}(t)) \end{cases}$$

Time-aware Cascade Model



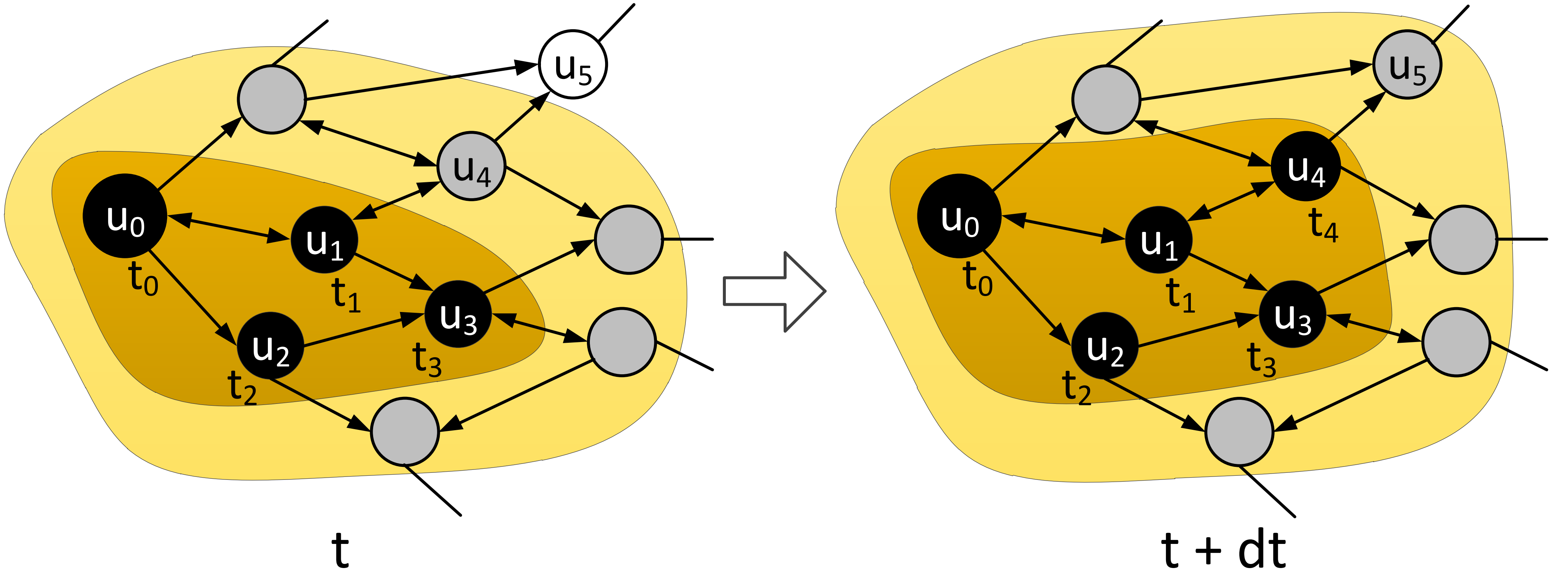
$$P_i(t) = h_i(t, \{t_j\}_{u_j \in Followee^{(i)}(t)}; \boldsymbol{\Theta}) \cdot dt$$

$$\begin{cases} P(\mathbb{C}(t+dt)) = P(\mathbb{C}(t+dt)|\mathbb{C}(t)) \cdot P(\mathbb{C}(t)) \\ P(\mathbb{C}(t_0)) = 1 \\ P(\mathbb{C}(t+dt)|\mathbb{C}(t)) = \boxed{\prod_{u_i \in \mathbb{X}^{(1)}(t)} P_i(t)} \cdot \prod_{u_{i'} \in \mathbb{X}^{(2)}(t)} (1 - P_{i'}(t)) \end{cases}$$

users who have re-shared

LARC
LIVING ANALYTICS RESEARCH CENTRE

Time-aware Cascade Model



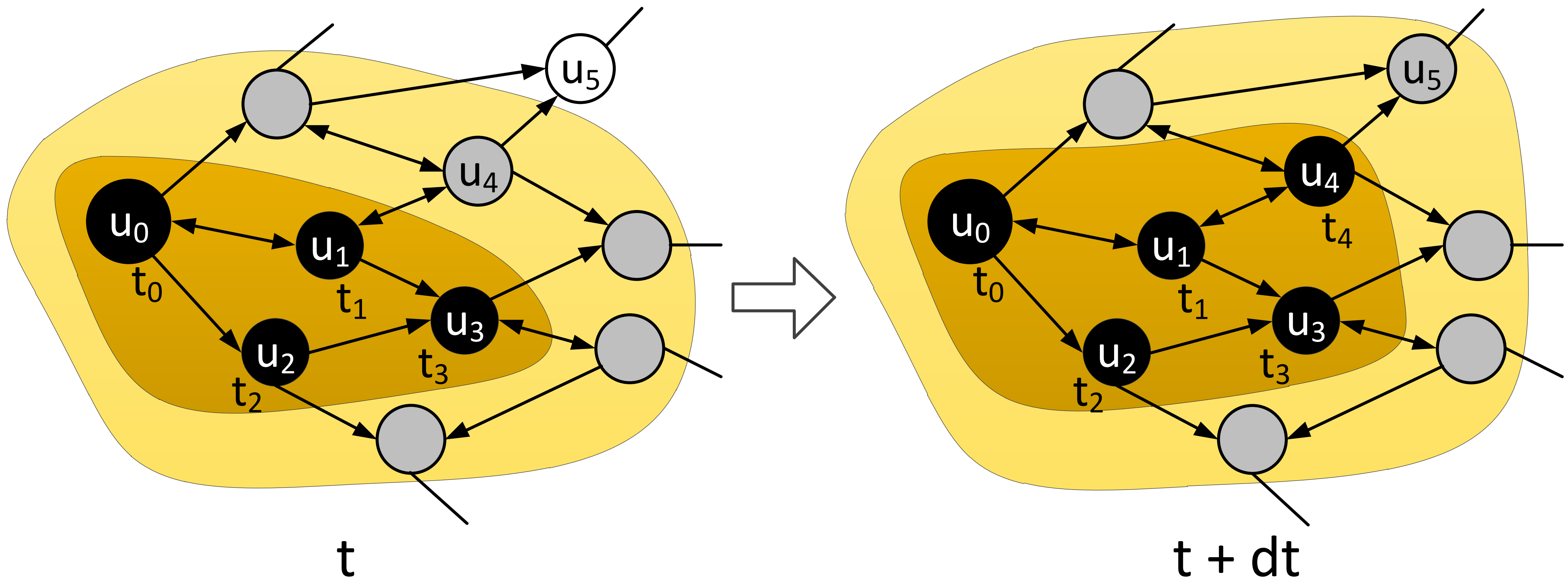
$$P_i(t) = h_i(t, \{t_j\}_{u_j \in Followee^{(i)}(t)}; \mathbf{\Theta}) \cdot dt$$

$$\begin{cases} P(\mathbb{C}(t+dt)) = P(\mathbb{C}(t+dt)|\mathbb{C}(t)) \cdot P(\mathbb{C}(t)) \\ P(\mathbb{C}(t_0)) = 1 \\ P(\mathbb{C}(t+dt)|\mathbb{C}(t)) = \prod_{u_i \in \mathbb{X}^{(1)}(t)} P_i(t) \cdot \prod_{u_{i'} \in \mathbb{X}^{(2)}(t)} (1 - P_{i'}(t)) \end{cases}$$

users who have re-shared users who haven't yet

LARC
LIVING ANALYTICS RESEARCH CENTRE

Time-aware Cascade Model



$$P_i(t) = h_i(t, \{t_j\}_{u_j \in Followee^{(i)}(t)}; \mathbf{\Theta}) \cdot dt$$

$$\begin{cases} P(\mathbb{C}(t+dt)) = P(\mathbb{C}(t+dt)|\mathbb{C}(t)) \cdot P(\mathbb{C}(t)) \\ P(\mathbb{C}(t_0)) = 1 \\ P(\mathbb{C}(t+dt)|\mathbb{C}(t)) = \prod_{u_i \in \mathbb{X}^{(1)}(t)} P_i(t) \cdot \prod_{u_{i'} \in \mathbb{X}^{(2)}(t)} (1 - P_{i'}(t)) \end{cases}$$

users who have re-shared　users who haven't yet

LARC
LIVING ANALYTICS RESEARCH CENTRE

Observations in Twitter

Observation 1. Only the first re-sharer matters.

$$P_i(t) = h_i(t, t_{j^\star}; \mathbf{\Theta}) \cdot dt$$

where $j^\star = argmin_j \{t_j | u_j \in Followee^{(i)}(t)\}$

LARC
LIVING ANALYTICS
RESEARCH CENTRE

Observations in Twitter

Observation 1. Only the first re-sharer matters.

$$P_i(t) = h_i(t, t_{j^\star}; \mathbf{\Theta}) \cdot dt$$

where $j^\star = argmin_j \{t_j | u_j \in Followee^{(i)}(t)\}$

Observation 2. The chance of a tweet to be retweeted decreases as time goes by.

$$P_i(t) = h_i(\tau; \mathbf{\Theta}) \cdot dt$$

where $\tau = t - t_{j^\star}$ and $h_i(\tau)$ is a **decreasing function**.

LARC
LIVING ANALYTICS RESEARCH CENTRE

Hazard Function Design

$$h(t) = \lim_{dt \to 0} \frac{P(t < T \leq t + dt | T > t)}{dt} = \frac{f(t)}{1 - F(t)}$$

Hazard Function Design

$$h(t) = \lim_{dt \to 0} \frac{P(t < T \leq t + dt | T > t)}{dt} = \frac{f(t)}{1 - F(t)}$$

$$H(t) = \int_0^t h(u) \mathrm{d}u = \int_0^t \frac{F'(u)}{1 - F(u)} \mathrm{d}u = -log(1 - F(u))|_0^t = -log(1 - F(t))$$

Hazard Function Design

$$h(t) = \lim_{dt \to 0} \frac{P(t < T \leq t + dt | T > t)}{dt} = \frac{f(t)}{1 - F(t)}$$

$$H(t) = \int_0^t h(u) \mathrm{d}u = \int_0^t \frac{F'(u)}{1 - F(u)} \mathrm{d}u = -log(1 - F(u))|_0^t = -log(1 - F(t))$$

$$F(t) = 1 - e^{-H(t)}$$

LARC
Living Analytics Research Centre

Hazard Function Design

$$h(t) = \lim_{dt \to 0} \frac{P(t < T \leq t + dt | T > t)}{dt} = \frac{f(t)}{1 - F(t)}$$

$$H(t) = \int_0^t h(u) \mathrm{d}u = \int_0^t \frac{F'(u)}{1 - F(u)} \mathrm{d}u = -log(1 - F(u))|_0^t = -log(1 - F(t))$$

$$F(t) = 1 - e^{-H(t)}$$

$$H(t) = \frac{t}{\lambda} \quad \Rightarrow \quad F(t) = 1 - e^{-\frac{t}{\lambda}}$$

Exponential distribution

LARC
LIVING ANALYTICS RESEARCH CENTRE

Hazard Function Design

$$h(t) = \lim_{dt \to 0} \frac{P(t < T \leq t + dt | T > t)}{dt} = \frac{f(t)}{1 - F(t)}$$

$$H(t) = \int_0^t h(u) \mathrm{d}u = \int_0^t \frac{F'(u)}{1 - F(u)} \mathrm{d}u = -log(1 - F(u))|_0^t = -log(1 - F(t))$$

$$F(t) = 1 - e^{-H(t)}$$

$H(t) = \frac{t}{\lambda} \quad \Rightarrow \quad F(t) = 1 - e^{-\frac{t}{\lambda}}$ **Exponential distribution**

$H(t) = (\frac{t}{\alpha})^{\beta} \quad \Rightarrow \quad F(t) = 1 - e^{-(\frac{t}{\alpha})^{\beta}}$ **Weibull distribution**

LARC
Living Analytics Research Centre

Hazard Function Design

$$H(t) = \frac{t}{\lambda} \quad \Rightarrow \quad F(t) = 1 - e^{-\frac{t}{\lambda}}$$

Exponential distribution

$$H(t) = (\frac{t}{\alpha})^{\beta} \quad \Rightarrow \quad F(t) = 1 - e^{-(\frac{t}{\alpha})^{\beta}}$$

Weibull distribution

LARC
LIVING ANALYTICS RESEARCH CENTRE

Hazard Function Design

$$H(t) = \frac{t}{\lambda} \quad \Longrightarrow \quad F(t) = 1 - e^{-\frac{t}{\lambda}}$$ **Exponential distribution**

$$H(t) = (\frac{t}{\alpha})^{\beta} \quad \Longrightarrow \quad F(t) = 1 - e^{-(\frac{t}{\alpha})^{\beta}}$$ **Weibull distribution**

$$H(\infty) = \infty \Rightarrow F(\infty) = 1 - e^{-\infty} \Rightarrow F(\infty) = 1$$

LARC
Living Analytics Research Centre

Hazard Function Design

$$H(t) = \frac{t}{\lambda} \quad \Rightarrow \quad F(t) = 1 - e^{-\frac{t}{\lambda}}$$ **Exponential distribution**

$$H(t) = (\frac{t}{\alpha})^{\beta} \quad \Rightarrow \quad F(t) = 1 - e^{-(\frac{t}{\alpha})^{\beta}}$$ **Weibull distribution**

$$H(\infty) = \infty \Rightarrow F(\infty) = 1 - e^{-\infty} \Rightarrow F(\infty) = 1$$

Hazard Function Design

$$H(t) = \frac{t}{\lambda} \quad \Rightarrow \quad F(t) = 1 - e^{-\frac{t}{\lambda}}$$ **Exponential distribution**

$$H(t) = (\frac{t}{\alpha})^{\beta} \quad \Rightarrow \quad F(t) = 1 - e^{-(\frac{t}{\alpha})^{\beta}}$$ **Weibull distribution**

~~$H(\infty) = \infty \Rightarrow F(\infty) = 1 - e^{-\infty} \Rightarrow F(\infty) = 1$~~

I) $H(0) = 0.$

II) $H(\infty) = -log(1 - F(\infty)) < \infty.$

III) $H(\tau)$ is an increasing function of τ.

IV) $h(\tau) = \frac{\mathrm{d}H(\tau)}{\mathrm{d}\tau}$ is a decreasing function of τ.

LARC
LIVING ANALYTICS RESEARCH CENTRE

Hazard Function Design

I) $H(0) = 0.$
II) $H(\infty) = -log(1 - F(\infty)) < \infty.$
III) $H(\tau)$ is an increasing function of τ.
IV) $h(\tau) = \frac{\mathrm{d}H(\tau)}{\mathrm{d}\tau}$ is a decreasing function of τ.

$$H(\tau) = \lambda \cdot (1 - (\frac{\tau}{\alpha} + 1)^{-\beta})$$

$$h(\tau) = \frac{\mathrm{d}H(\tau)}{\mathrm{d}\tau} = \lambda \cdot \frac{\beta}{\alpha} \cdot (\frac{\tau}{\alpha} + 1)^{-(\beta+1)}$$

LARC
LIVING ANALYTICS RESEARCH CENTRE

Hazard Function Design

$$H(\tau) = \lambda \cdot (1 - (\frac{\tau}{\alpha} + 1)^{-\beta})$$

LARC
LIVING ANALYTICS
RESEARCH CENTRE

Hazard Function Design

$$H(\tau) = \lambda \cdot (1 - (\frac{\tau}{\alpha} + 1)^{-\beta})$$

scale parameter

LARC
LIVING ANALYTICS RESEARCH CENTRE

Hazard Function Design

$$H(\tau) = \lambda \cdot (1 - (\frac{\tau}{\alpha} + 1)^{-\beta})$$

scale parameter

shape parameter

LARC
LIVING ANALYTICS
RESEARCH CENTRE

Hazard Function Design

$$H(\tau) = \lambda \cdot (1 - (\frac{\tau}{\alpha} + 1)^{-\beta})$$

scale parameter

shape parameter

$$F(\infty) \approx H(\infty) = \lambda$$

Hazard Function Design

$$H(\tau) = \lambda \cdot (1 - (\frac{\tau}{\alpha} + 1)^{-\beta})$$

scale parameter

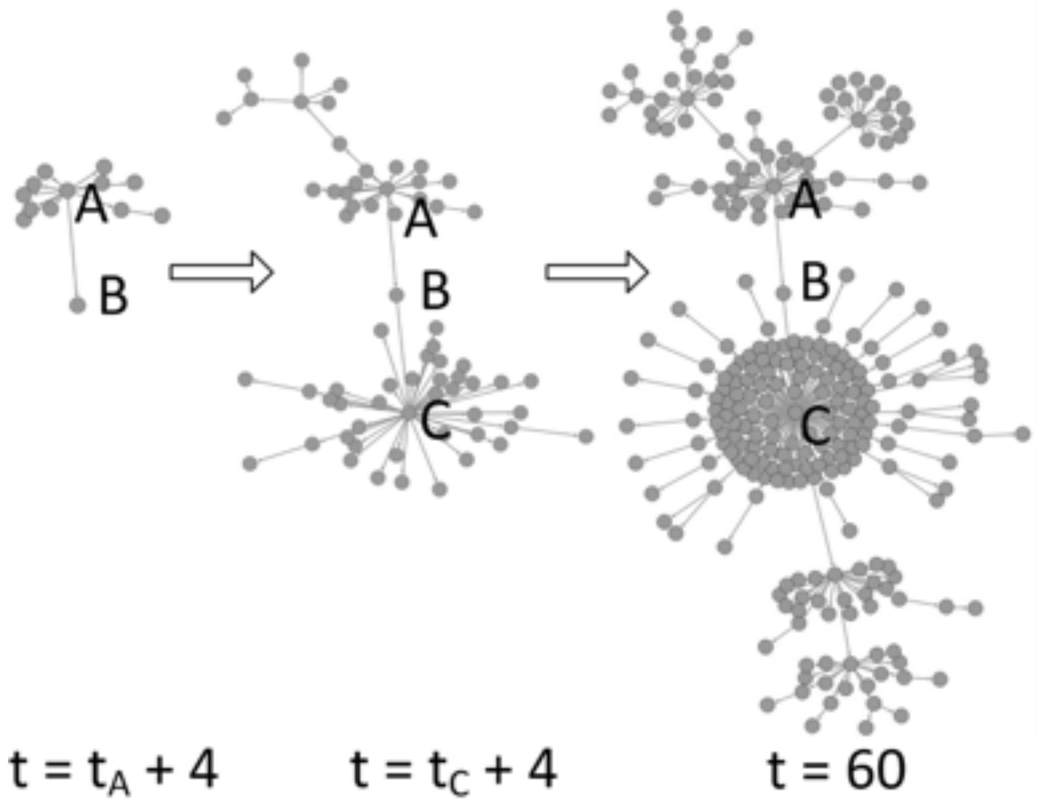
shape parameter

$$F(\infty) \approx H(\infty) = \lambda$$

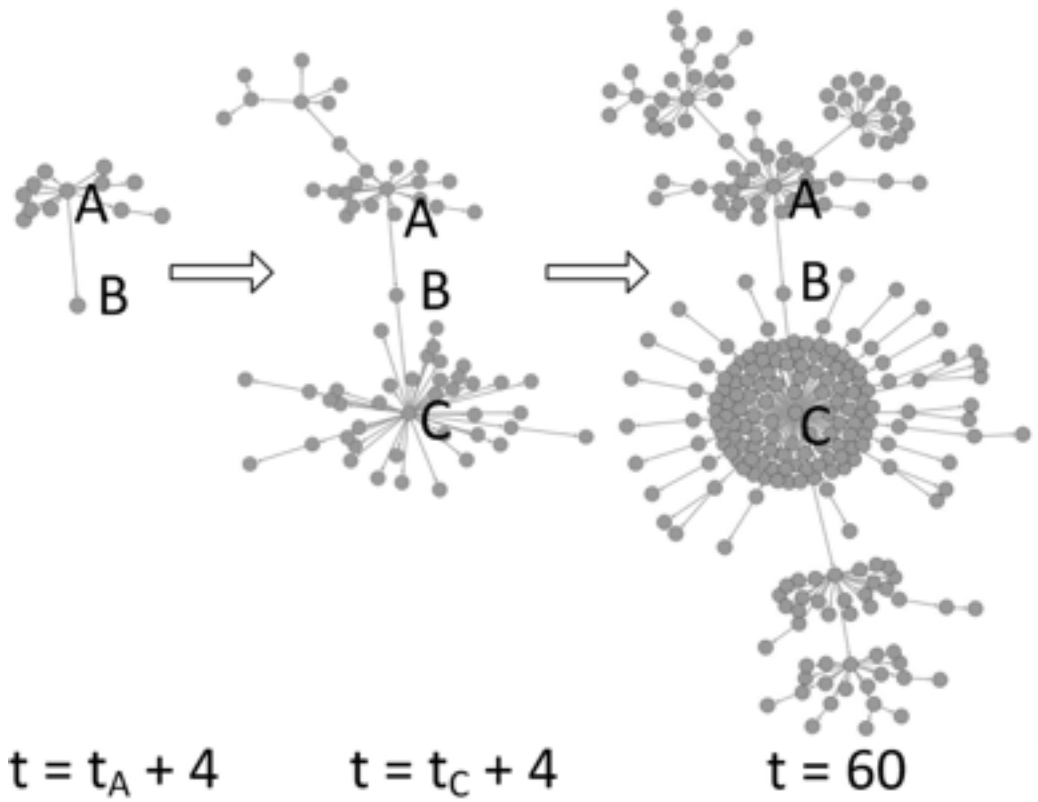
describes the eventual re-tweeting probability

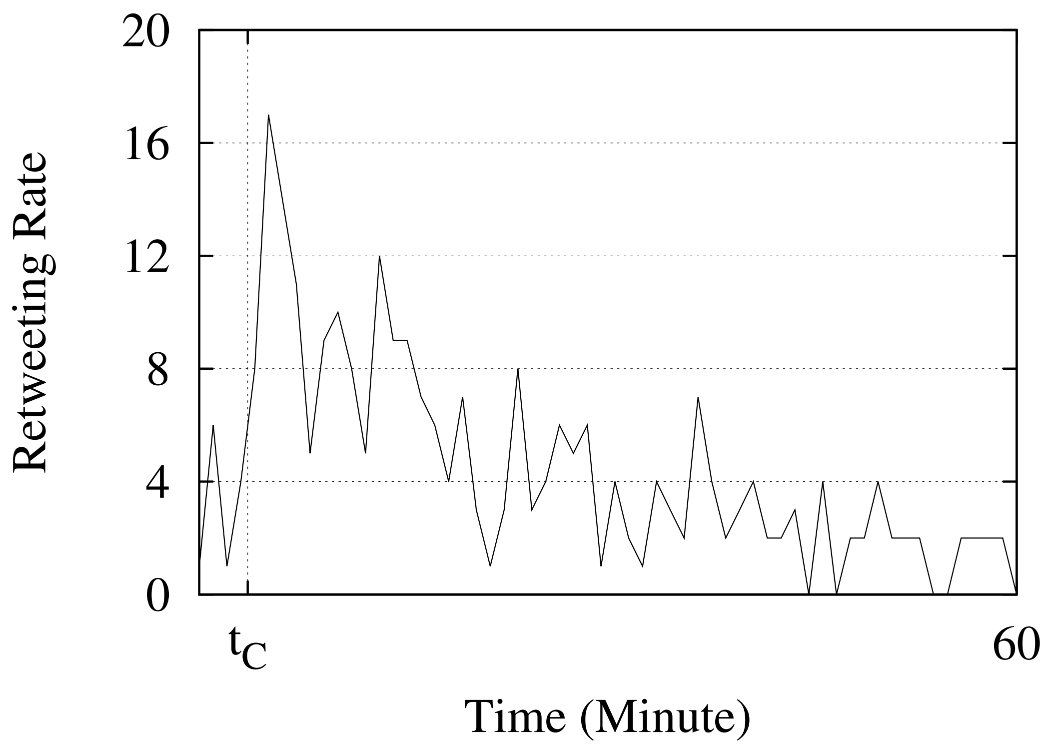
LARC
LIVING ANALYTICS RESEARCH CENTRE

Hazard Rate Illustration

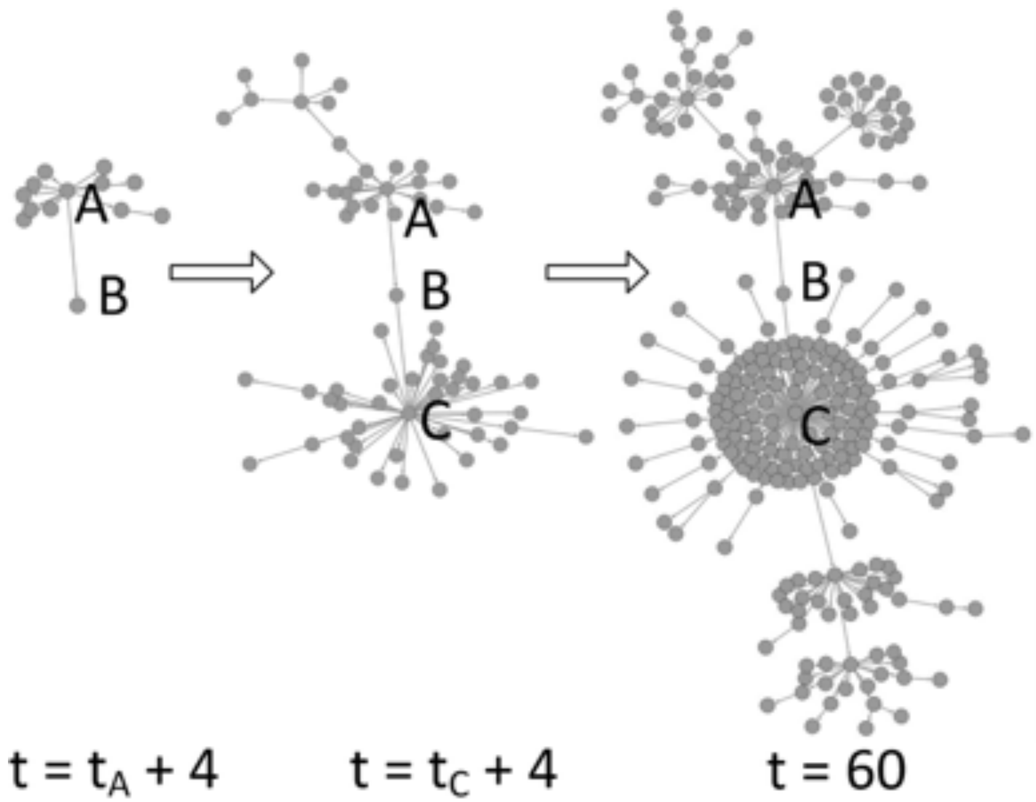


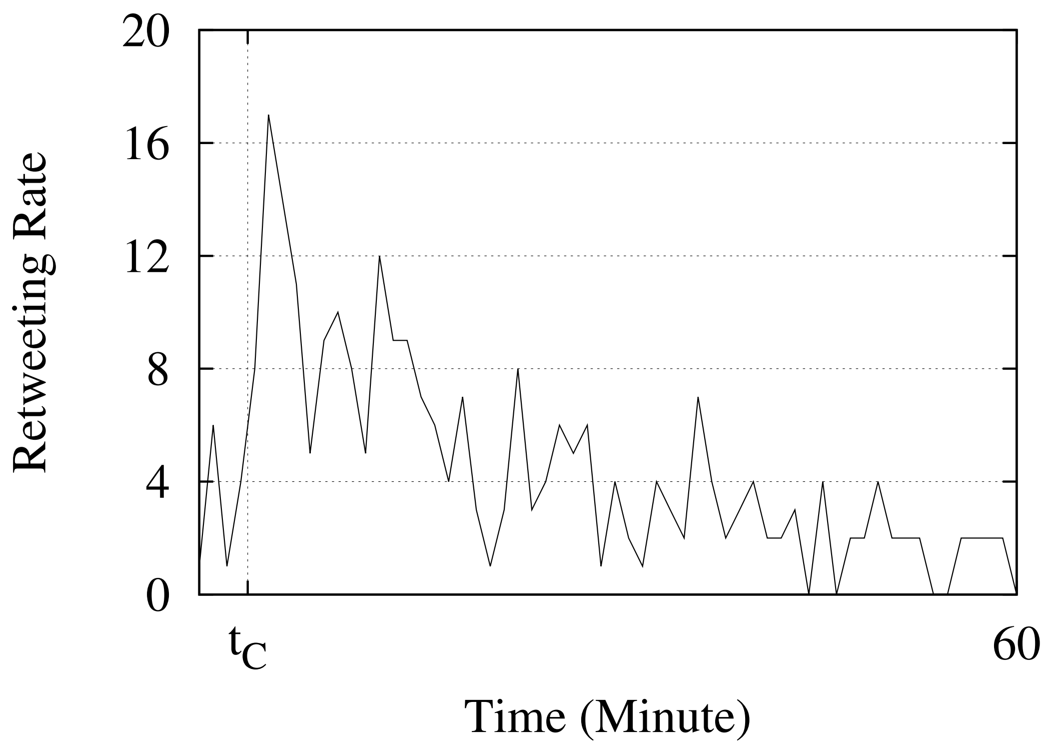
Hazard Rate Illustration

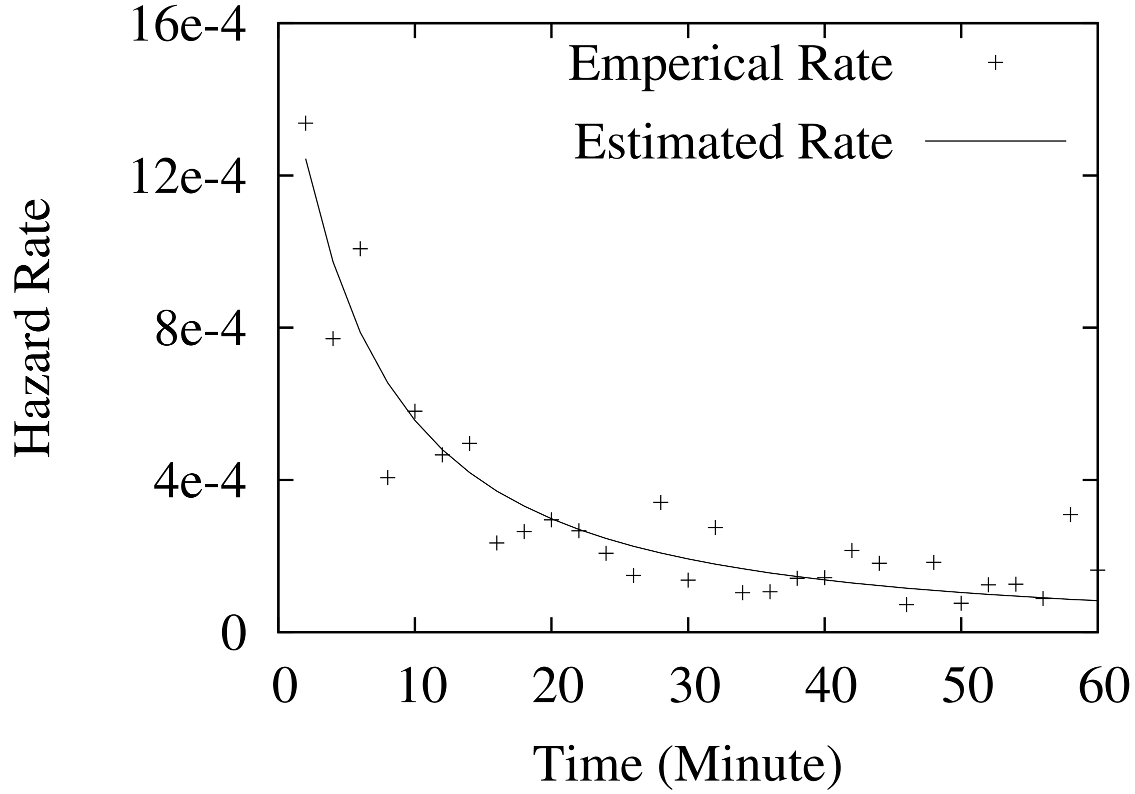




Hazard Rate Illustration

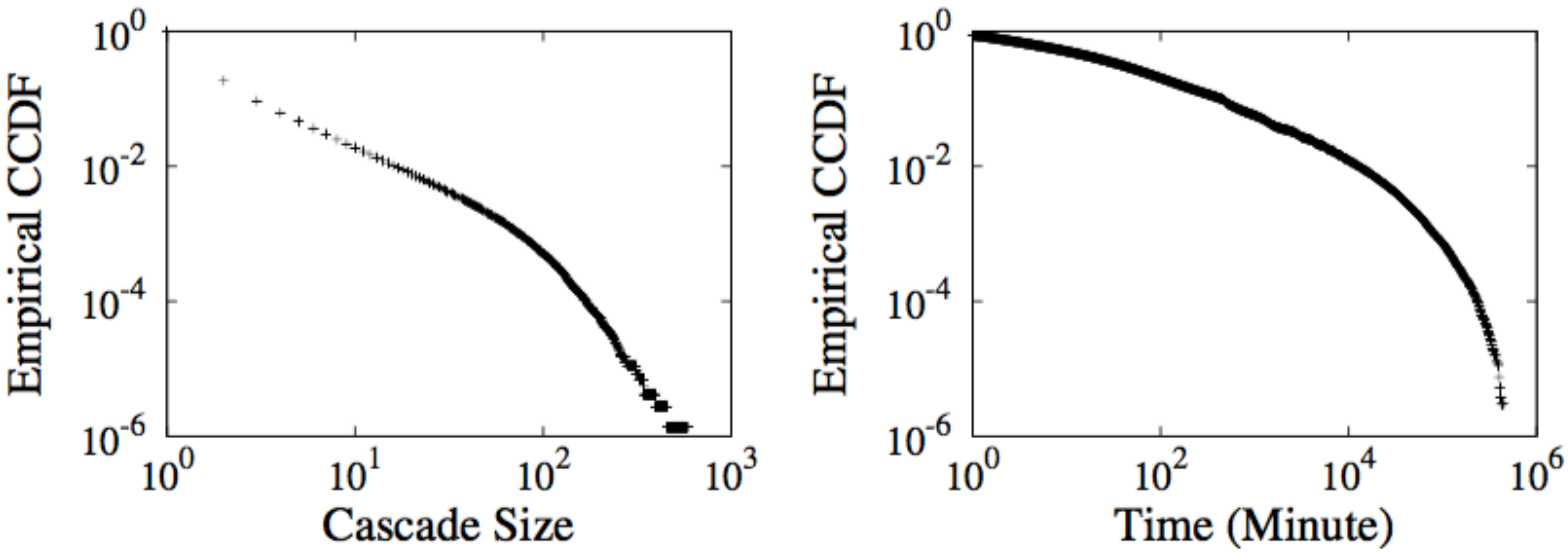






Dataset

From a Singapore based Twitter data set, we get all the retweets to construct retweeting cascades. In all we get 2,425,348 cascades.



LARC
LIVING ANALYTICS RESEARCH CENTRE

Probabilistic Model Fitting

- **TM_t** Threshold Model

$$h_i(t) = \lambda \cdot s(|Followee^{(i)}(t)|)$$

where $s(x) = \frac{1}{1 + e^{-a(x-b)}}$

- **TCM-CH** Constant Hazard

$$H(\tau) = \lambda \cdot \tau \qquad h(\tau) = \frac{\mathrm{d}H(\tau)}{\mathrm{d}\tau} = \lambda$$

- **TCM-EH** Exponential Hazard

$$H(\tau) = \lambda \cdot (1 - e^{-k \cdot \tau}) \qquad h(\tau) = \frac{\mathrm{d}H(\tau)}{\mathrm{d}\tau} = \lambda \cdot k \cdot e^{-k \cdot \tau}$$

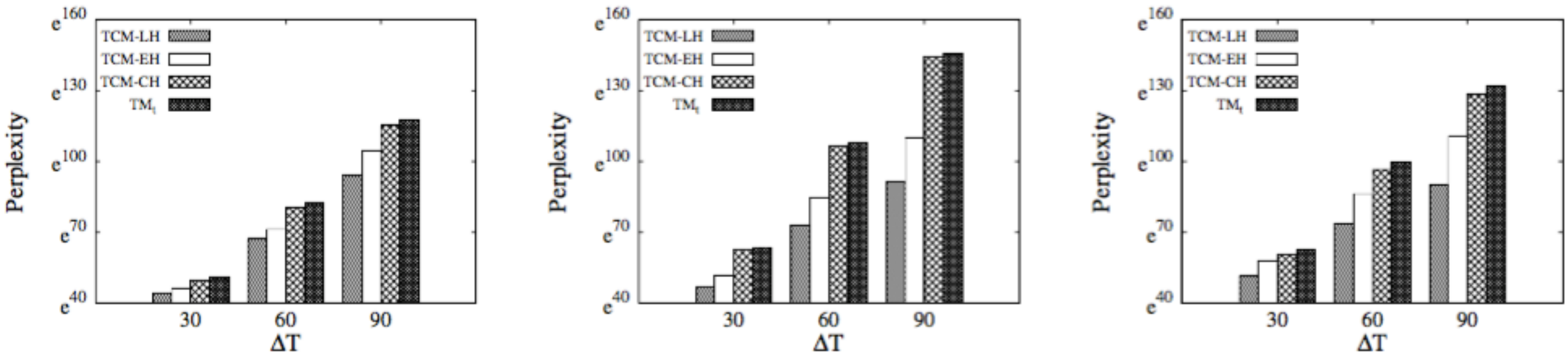
- **TCM-LH** Long tail Hazard (our proposed)

$$H(\tau) = \lambda \cdot (1 - (\frac{\tau}{\alpha} + 1)^{-\beta}) \qquad h(\tau) = \frac{\mathrm{d}H(\tau)}{\mathrm{d}\tau} = \lambda \cdot \frac{\beta}{\alpha} \cdot (\frac{\tau}{\alpha} + 1)^{-(\beta+1)}$$

LARC
LIVING ANALYTICS RESEARCH CENTRE

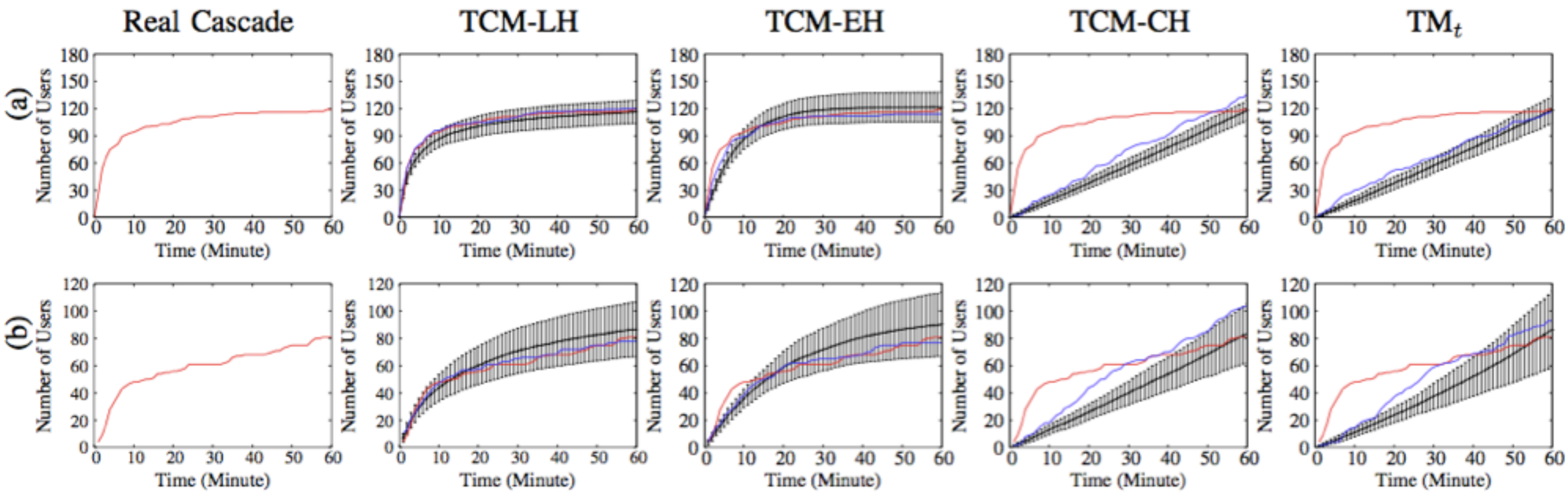
Probabilistic Model Fitting

For each cascade, observe its development in first T_0 for training, and the next ΔT for testing.

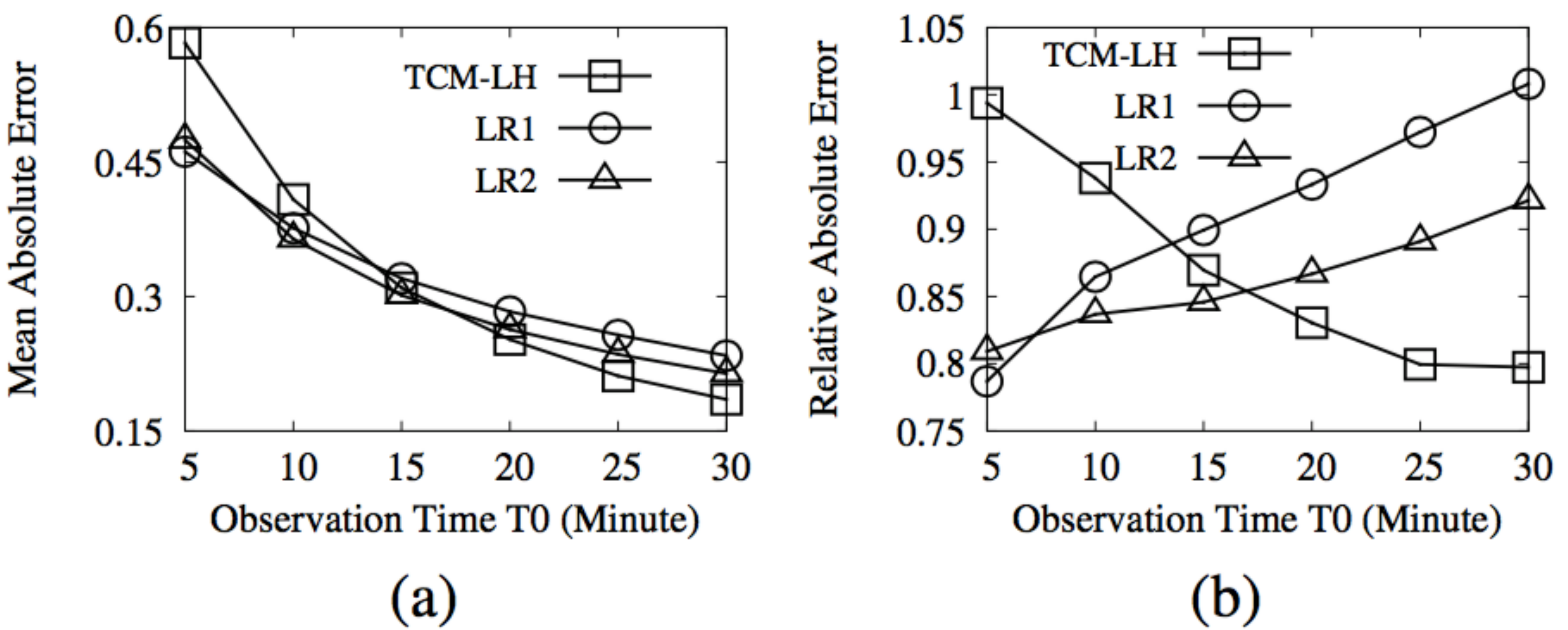


(a) Year 2010 (b) Year 2011 (c) Year 2012

Probabilistic Model Fitting



Predicting Cascade Growth



LARC
LIVING ANALYTICS RESEARCH CENTRE

Virality Prediction

Threshold	Measure	Random Guessing	Without Simulation	With Simulation
20	Recall	0.4817	0.4535	**0.6254**
	Precision	0.0034	**0.7285**	0.5678
	F1	0.0068	0.5590	**0.5952**
25	Recall	0.5764	0.4716	**0.5808**
	Precision	0.0026	**0.7500**	0.6215
	F1	0.0053	0.5791	**0.6005**
30	Recall	0.4600	0.4333	**0.5667**
	Precision	0.0014	**0.6915**	0.6071
	F1	0.0027	0.5328	**0.5862**
35	Recall	0.4653	0.3762	**0.5446**
	Precision	0.0009	**0.6909**	0.5612
	F1	0.0019	0.4872	**0.5528**
40	Recall	0.4545	0.2424	**0.4697**
	Precision	0.0006	**0.6667**	0.4247
	F1	0.0012	0.3556	**0.4460**

LARC
Living Analytics Research Centre

Thanks

Our work is based on previous cascade models

- J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of- mouth. Marketing letters, 12(3):211–223, 2001.

- M.Gomez-Rodriguez,D.Balduzzi,andB.Scho¨lkopf.Uncovering the temporal dynamics of diffusion networks. In Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011, pages 561–568, 2011.

- S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In The 18th ACM SIGKDD Inter- national Conference on Knowledge Discovery and Data Mining, KDD ’12, Beijing, China, August 12-16, 2012, pages 33–41, 2012.

- M. Gomez-Rodriguez, J. Leskovec, and B. Scho¨lkopf. Modeling information propagation with survival theory. In ICML (3), pages 666–674, 2013.

- N. Du, L. Song, M. Gomez-Rodriguez, and H. Zha. Scalable influence estimation in continuous-time diffusion networks. In Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States., pages 3147–3155, 2013.