Elastic and Realistic Social Media Data Generation

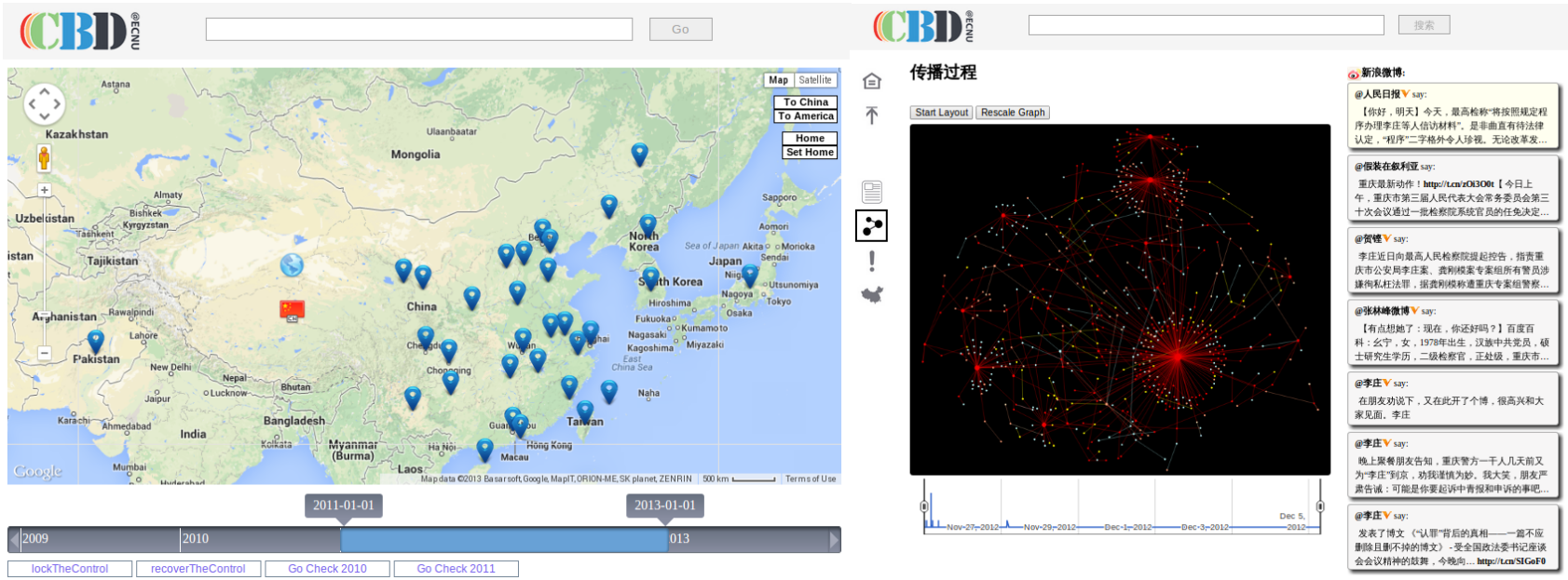
Weining Qian, Minqi Zhou

Center for Cloud Computing and Big Data
East China Normal University
{wnqian,mqzhou}@sei.ecnu.edu.cn

November 19, 2013

Outline

Social media as a source for collective behavior analysis



http://database.ecnu.edu.cn/microblogcube/index_debug.html

Motivation: A benchmark is needed

The demonstration system

- Based on data crawled from Sina Weibo via API
  - Sina Weibo: A Chinese Twitter-like social media service
  - All tweets of about 2 million users are used (Oct. 2009 - Jun. 2013)
  - 200 hotspots are annotated
- Analytics focus on content and *network patterns*
  - Spamming and marketing/advertising behavior identification
  - Modeling of hotspot evolution
  - Hotspot monitoring and prediction

To benchmark mgmt. and mining technologies for social media data

- Efficiently generate realisic data (this talk)
- Analytical queries
- Measurements and performance testing tools

Outline

BSMA-GEN: Generating tweet timeline structures

A tweet is a tuple $< t, c, u, f >$

- t the timestamp when the tweet is published
- c the content of the tweet
- u the author
- f is a pointer, point to the father of the tweet
  - *null* for original tweet, $m \rightarrow n$ for retweets.

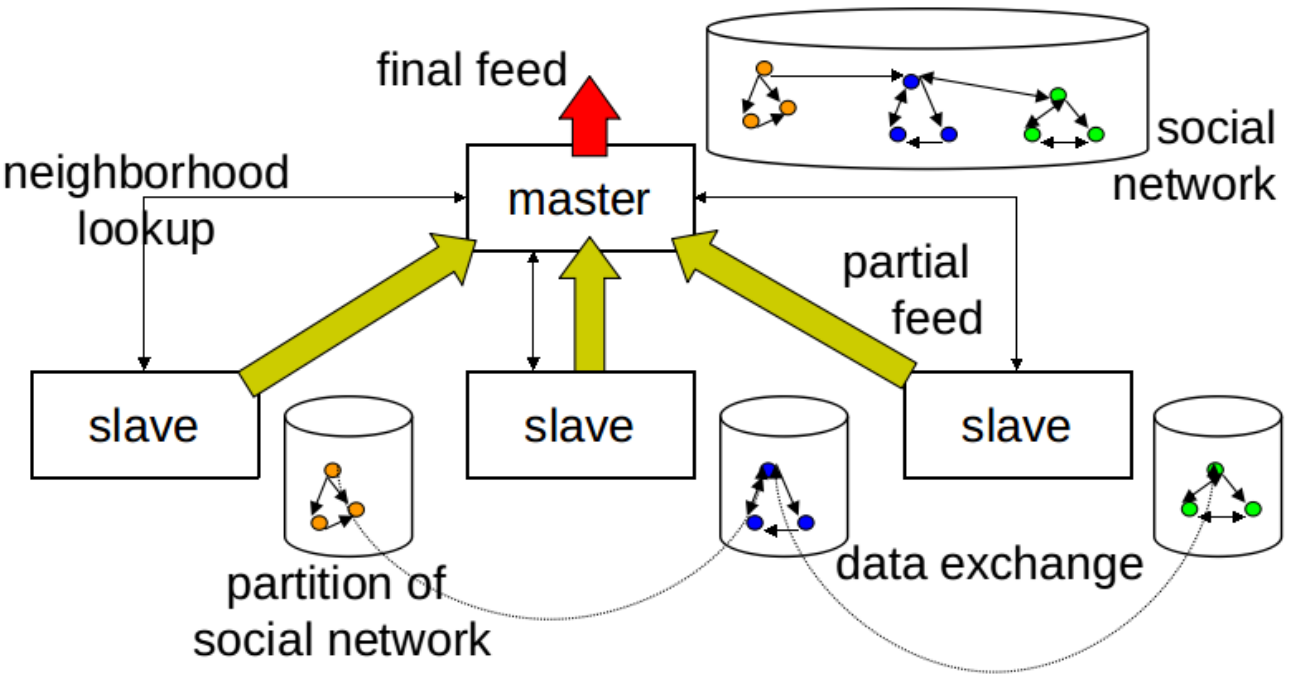
Requirements

Data should preserve distributions

- Degree distributions of the followship network
  - Solved by several previous work
- Retweet frequency distribution over users and tweets
- Tweet and retweet interval distribution of users and global timeline

Outline

Framework of BSMA-GEN



Tweet generation

Model

User i publish tweets $N(t,i), t \geq 0$ can be modeled as a Nonhomogeneous Poisson Process with changing intensity function $\lambda_i(t)$:

$$\begin{aligned} \lambda_i(t) &= \lambda_i \cdot f(t) \\ f(t) &= D_t \cdot H_t \end{aligned}$$

Generation

Thinning algorithm Nonhomogeneous Poisson Process for each user

$NextTime(i,t)$ At time t, the next timestamp for user i to tweet is determined.

Retweet generation

1. Determine if a tweet is a retweet
2. For each retweet, determine its parent tweet

Social network generation

Edge copying model is used

Streaming the generation process

Data strutures

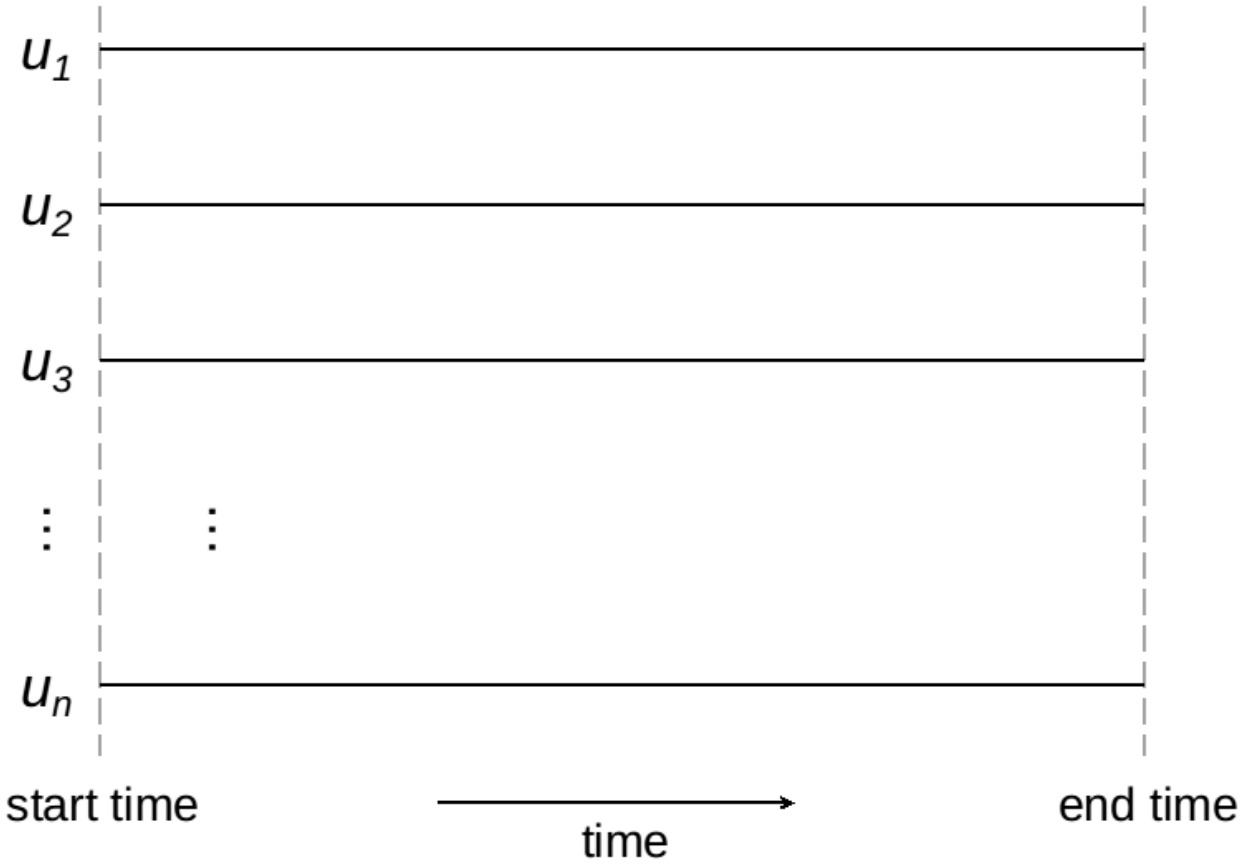
Following pool Keep the next tweet time of each user

Candidate pool Keep tweets in the window size

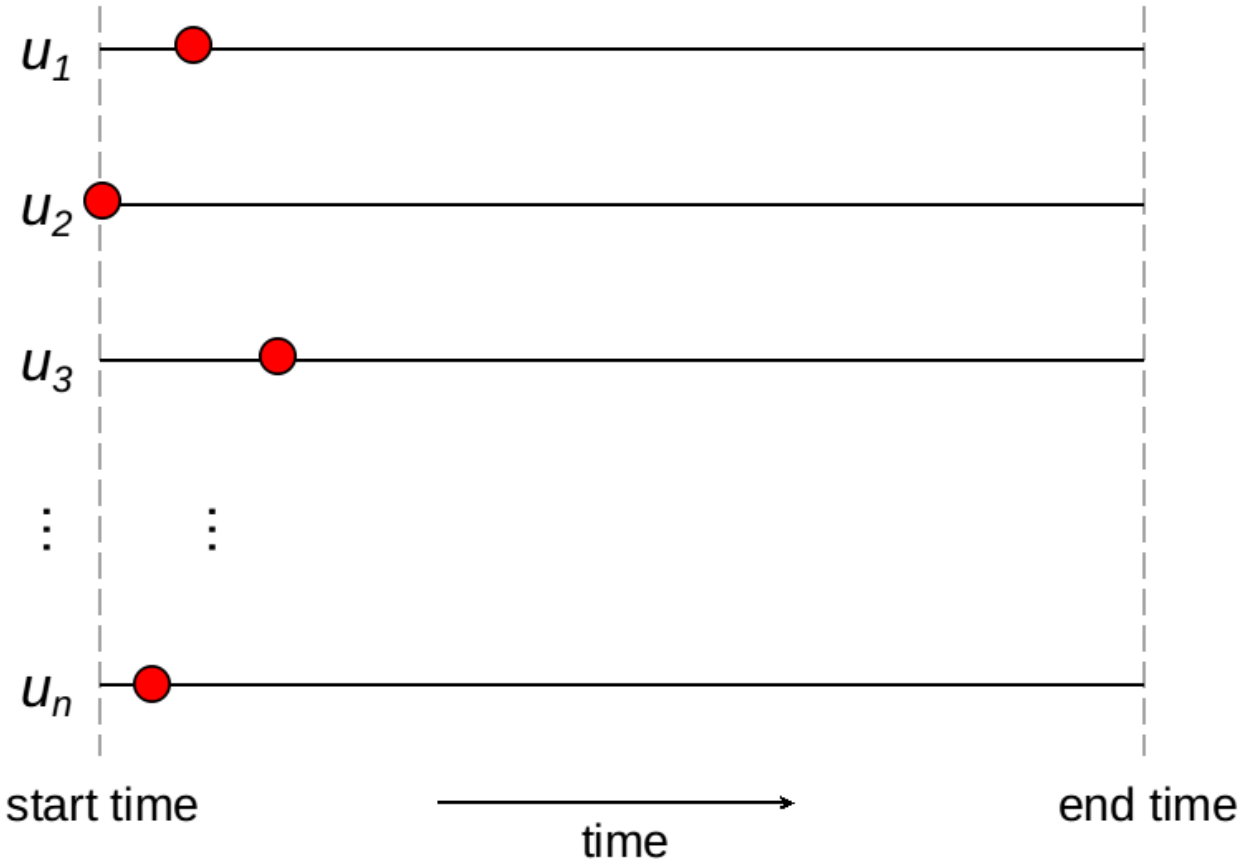
Process

1. Initialize the following pool
2. Update two pools
   - Move tweets from the following pool to candidate pool in chronological order

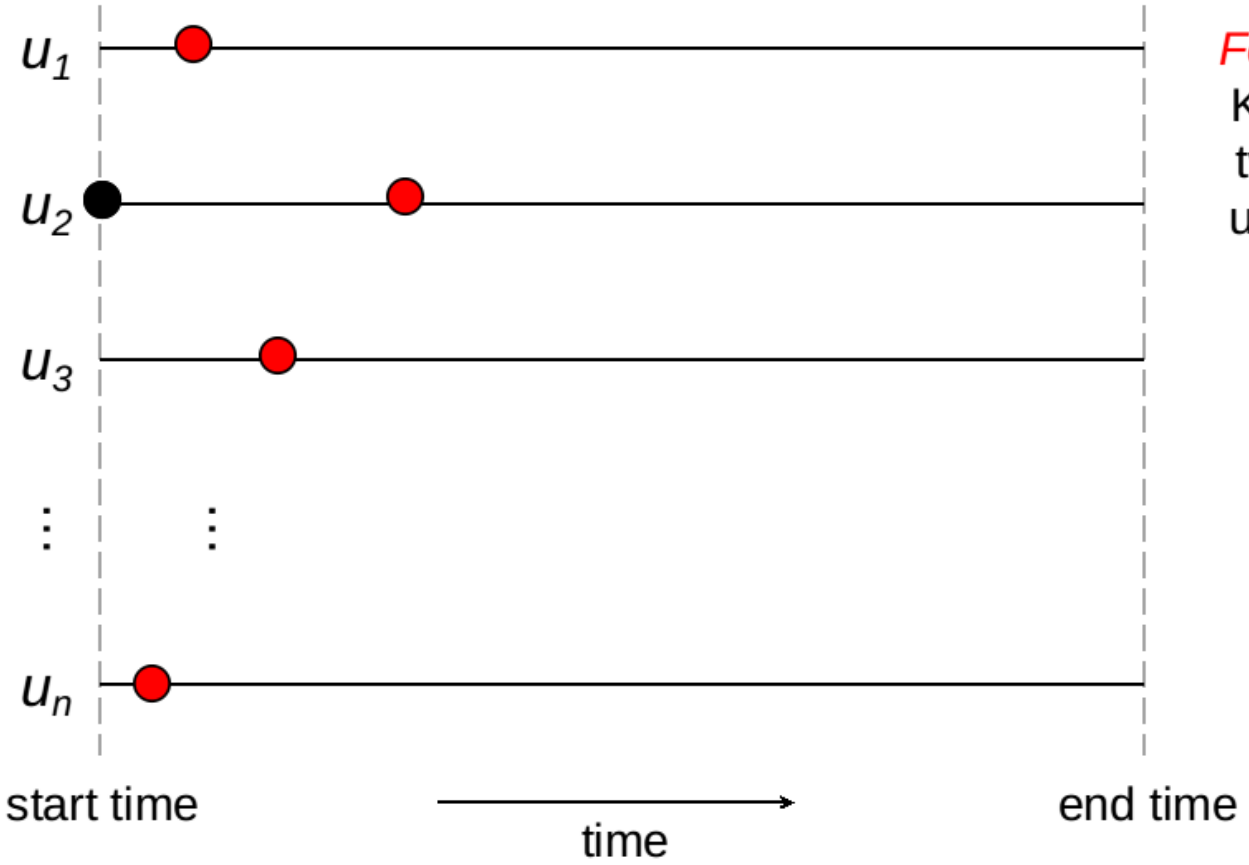
Streaming the generation process



Streaming the generation process

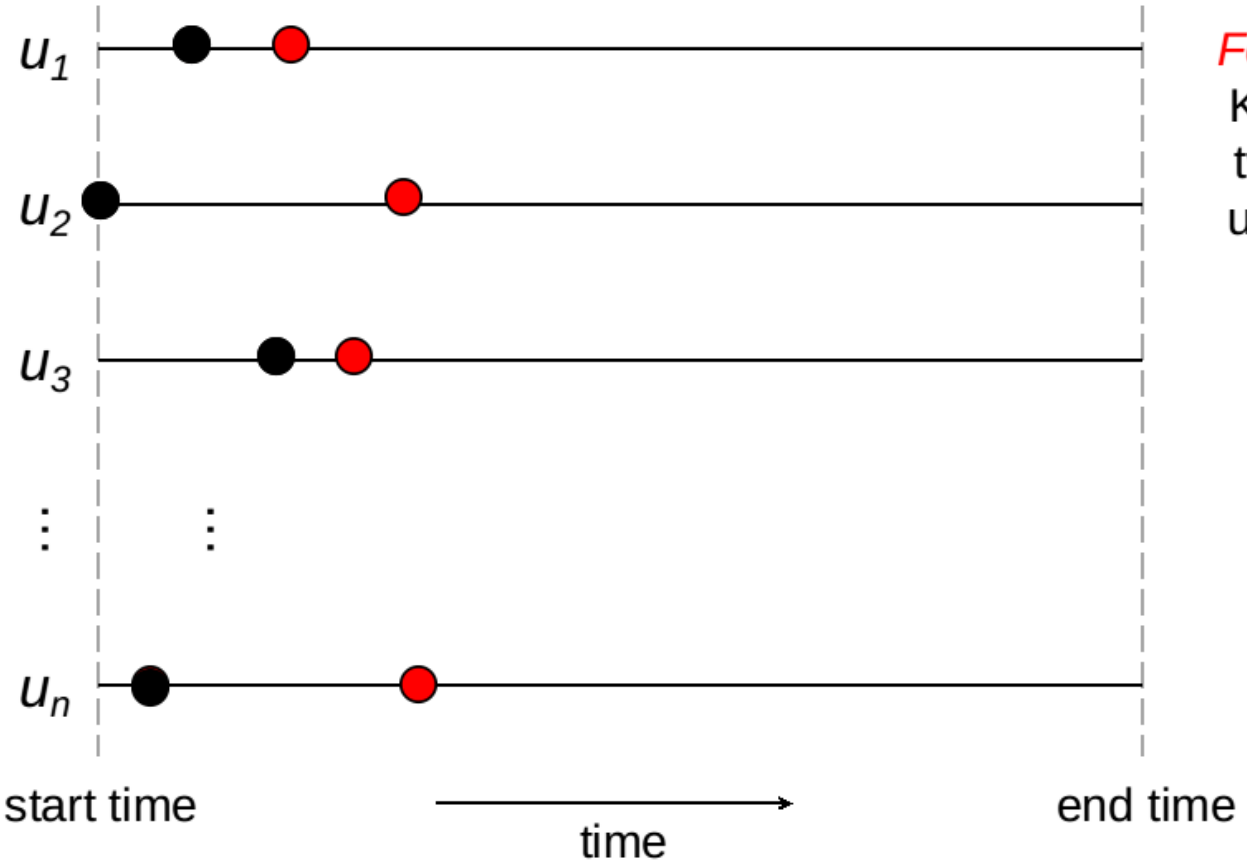


Streaming the generation process



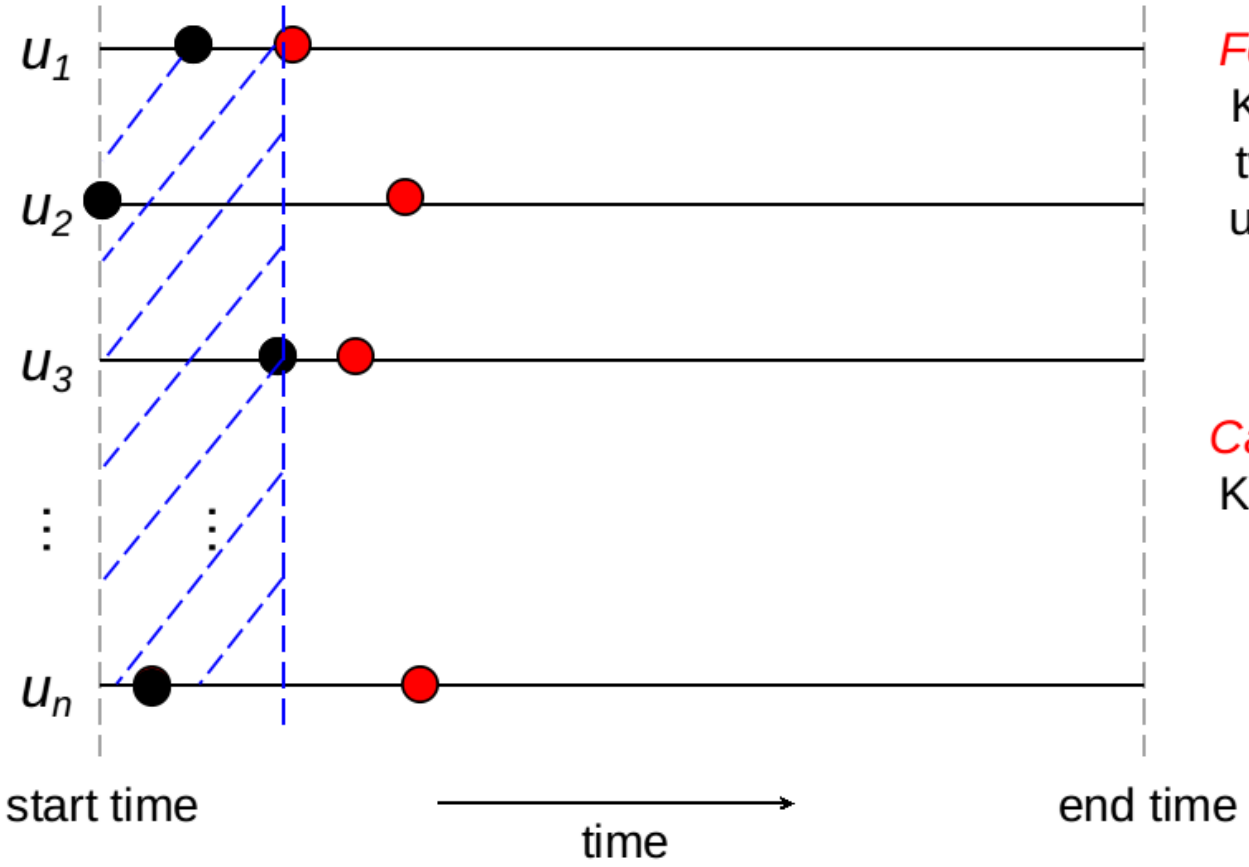
Following pool:
Keep the next tweet of each user (red one)

Streaming the generation process



Following pool:
Keep the next tweet of each user (red one)

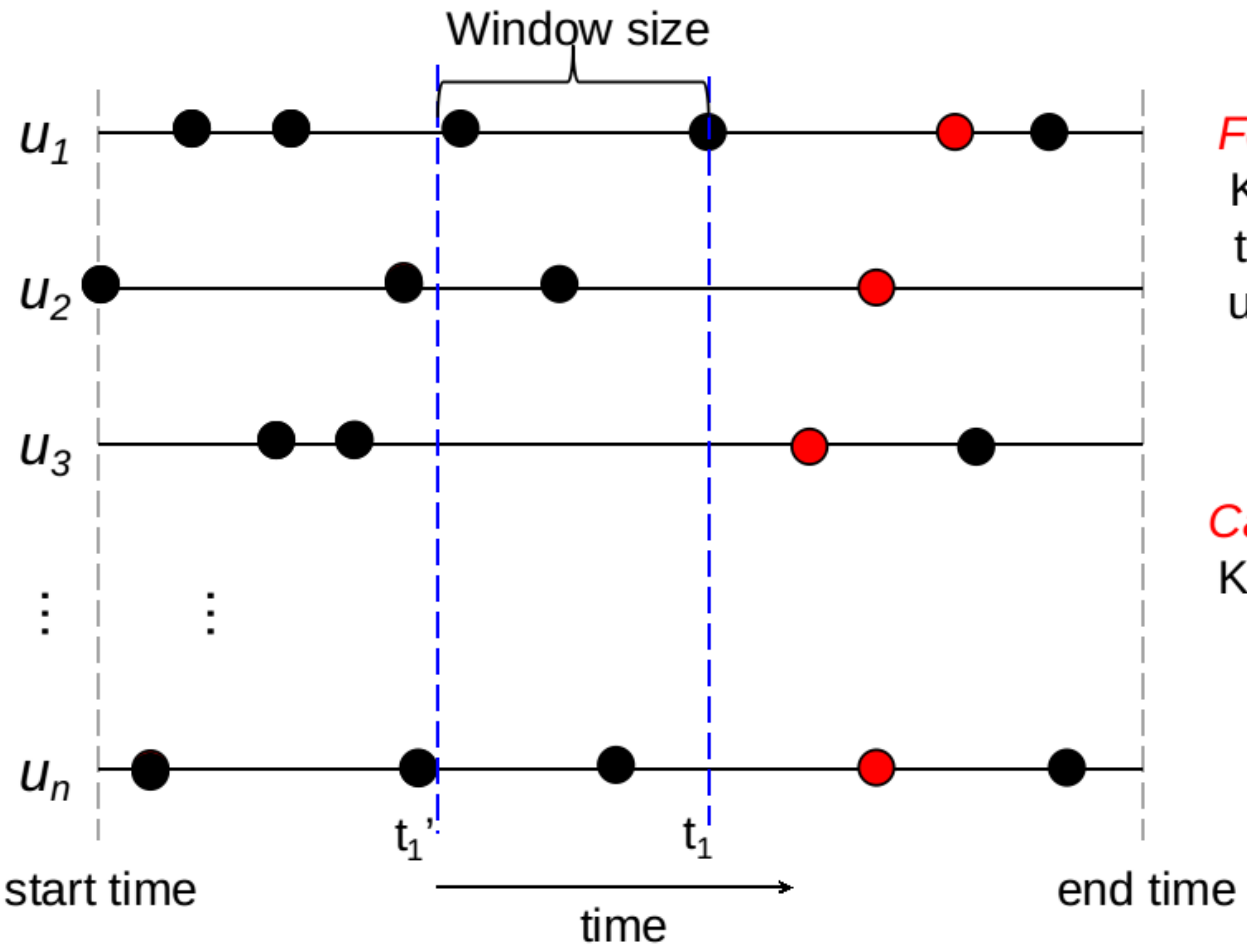
Streaming the generation process



Following pool:
Keep the next tweet of each user (red one)

Candidate pool:
Keep tweets in the window size(*W*)

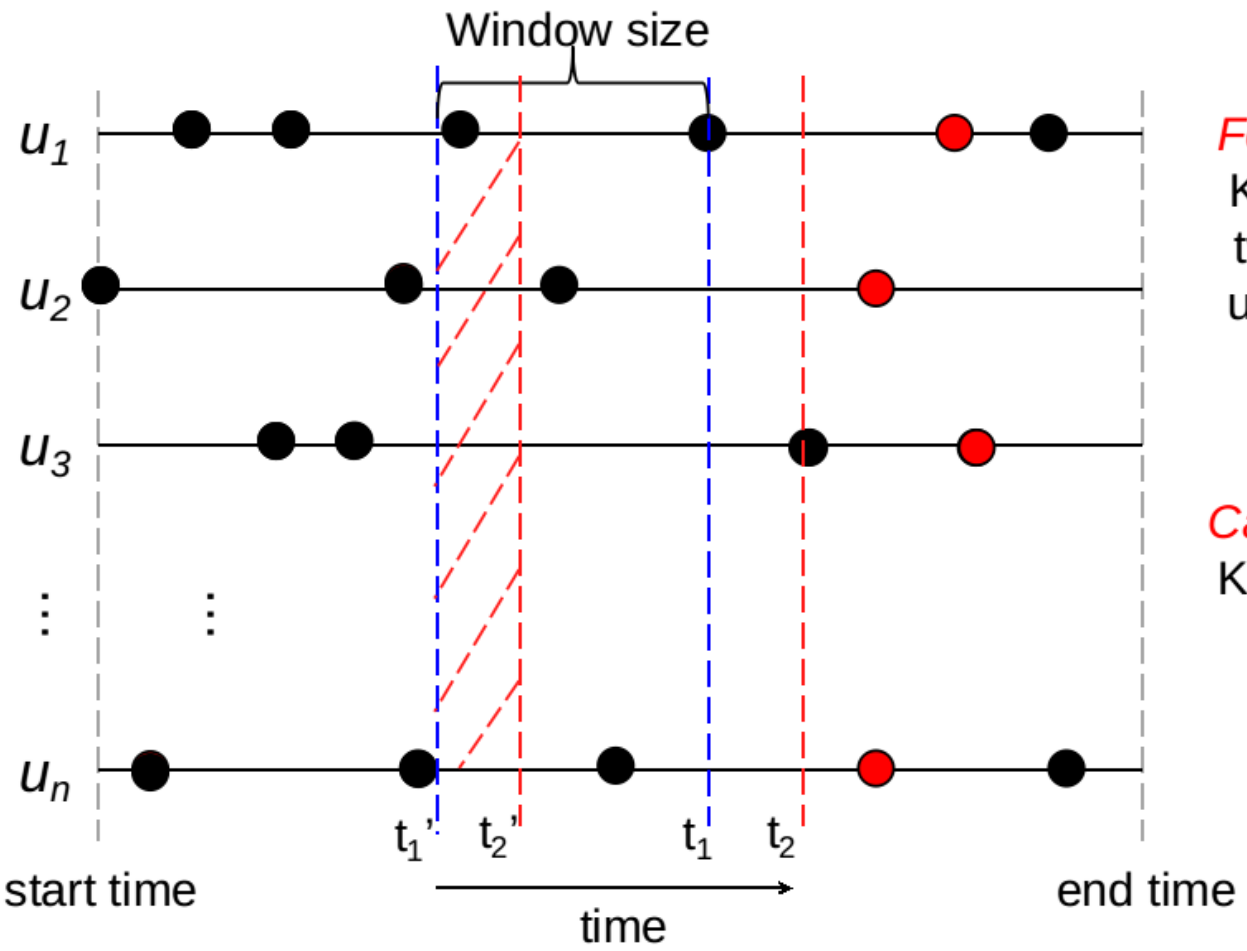
Streaming the generation process



Following pool: Keep the next tweet of each user (red one)

Candidate pool: Keep tweets in the window size(*W*)

Streaming the generation process



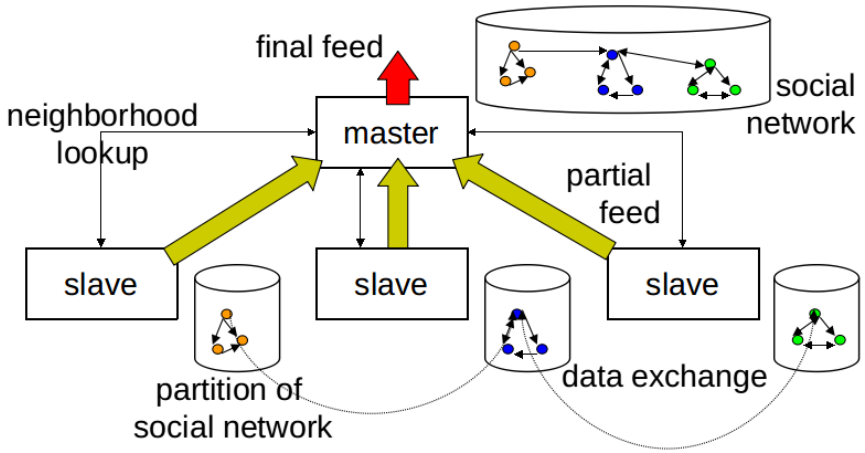
Following pool: Keep the next tweet of each user (red one)

Candidate pool: Keep tweets in the window size(*W*)

Outline

Parallel generation

- Master is in responsible for partition the social network and assign tasks
- Each slave is in responsible for generating tweets of users in ints partition
- Communication is triggered when interaction is needed
  - *Asynchronized communication* and *delayed retweet publishing*
- Final timeline is merged by the master

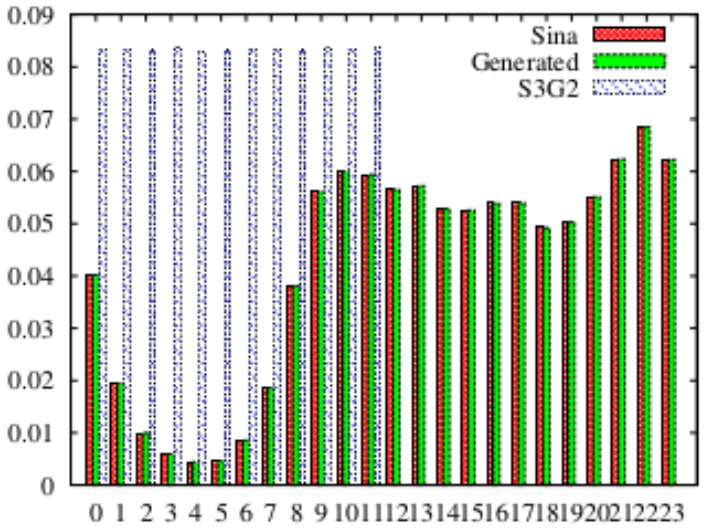


Outline

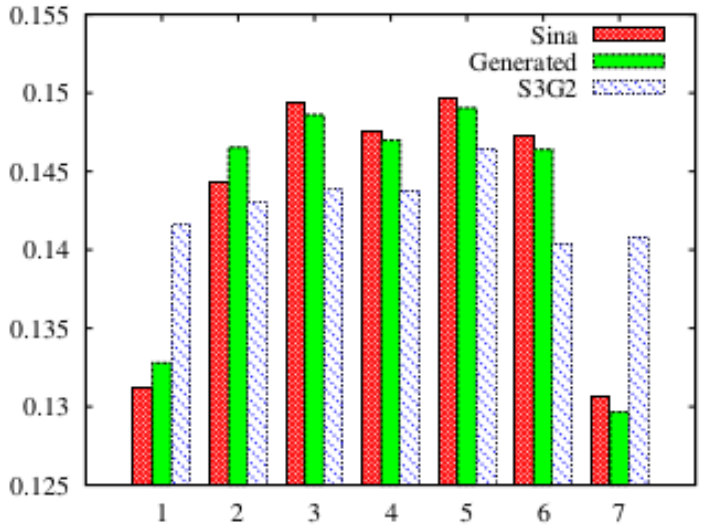
Experimental setup

- 100,000 to 1,000,000 users
- 5 nodes cluster (1 master and 4 slaves)
- To generate a 1 year timeline
- All parameters are learned automatically from the Sina Weibo data

Distribution of user activity over time

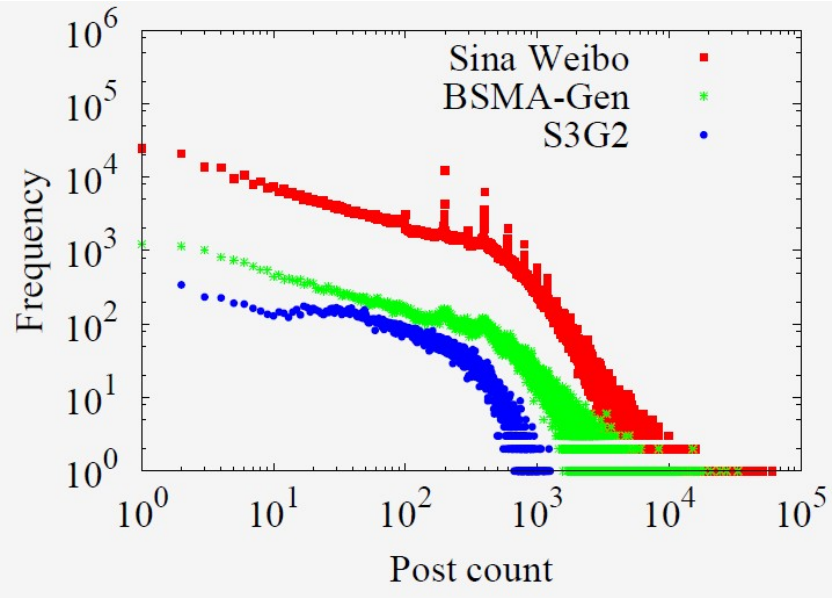


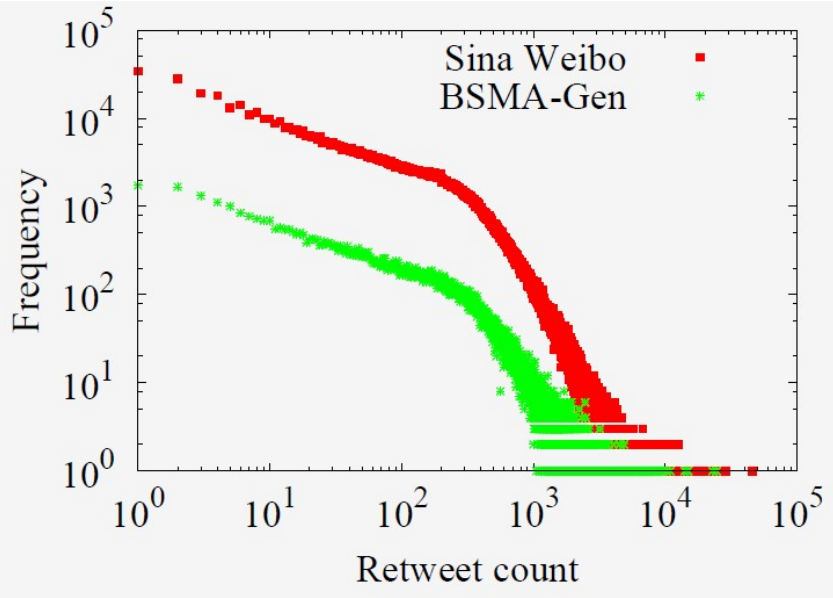
(a) Distribution of activities over hour



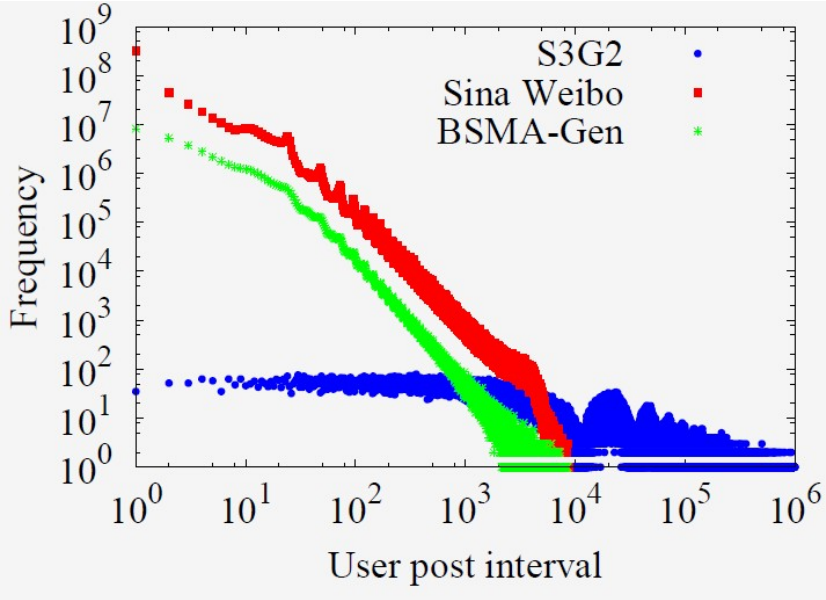
(b) Distribution of activities over day

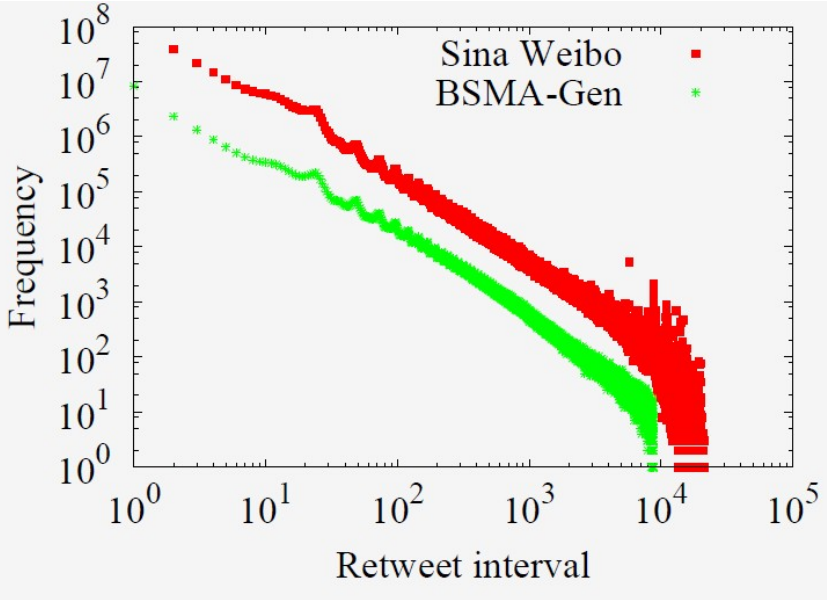
Distribution of #tweet and #retweet



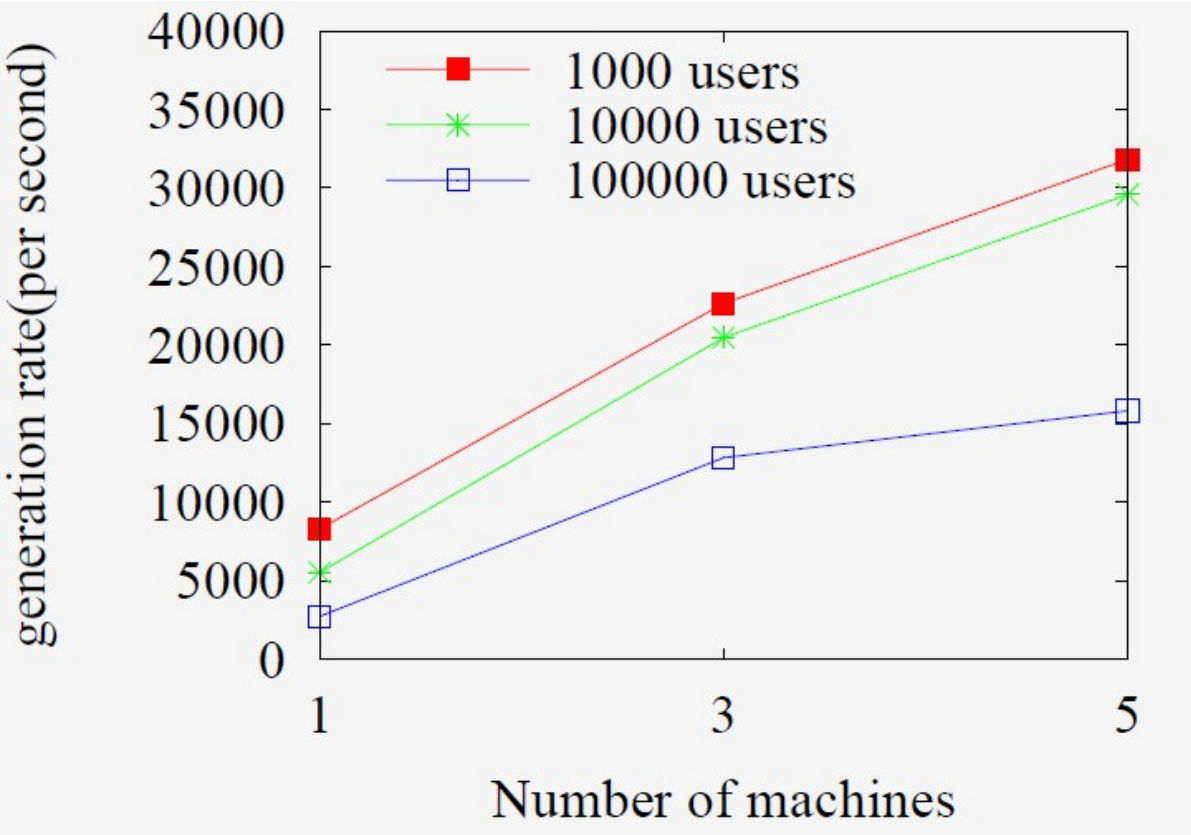


Distribution of user activity intervals





Speedup


40000
35000
30000
25000
20000
15000
10000
5000
0
generation rate(per second)
1000 users
10000 users
100000 users
1
3
5
Number of machines

Outline

Summary

- BSMA-GEN is designed to generate realistic social media data, for benchmarking purpose
  - https://github.com/c3bd/BSMA
  - Specifically that are *similar* to Sina Weibo data
- Future work/requirements include:
  - More efficient parallel process
    - The bottleneck is in the access to the followship network
  - To simulate timelines of other social media data
  - To generate event tagging, (Chinese) content, etc.
  - As a complementary to other benchmarks, e.g. SNB/LDBC

Thanks!

http://database.ecnu.edu.cn/