Aggregate Estimation over a Microblog Platform

Saravanan Thirumuruganathan	UTA
Nan Zhang	GWU
Vagelis Hristidis	UCR
Gautam Das	UTA

Outline

- Motivation
- Challenges
- Subgraph Design
- Sampling Design
- Experiments

Motivation

Microblog Platforms

twitter

yammer

tumblr.

Instagram

Twitter has gone Mainstream

Aggregate Estimation over Microblogs

Aggregates	Estimands	Filters
AVERAGE	User Profile metadata	Keywords
COUNT	Tweets metadata	Time
SUM		

Motivation - I

Sociologist:

- COUNT(privacy) before and after Snowden disclosures
- COUNT(snowden, hero) Vs COUNT(snowden, traitor)

Motivation - II

Psychologist:

- What fraction of Twitter users have tweets with keywords symptomatic of depression?

Motivation - III

Companies/Investors:

- COUNT(company Name) Month-over-Month in 2013
- COUNT(android) Vs COUNT(iPhone) in 2013

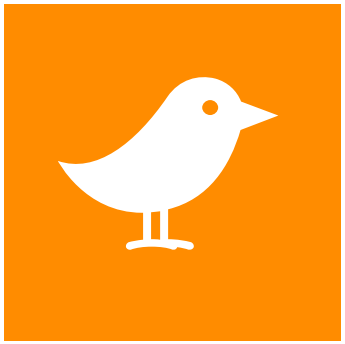
Access Model

SEARCH

Given a keyword, returns recent tweets containing keyword

USER CONNECTIONS

Given a user, returns his/her connections



USER TIMELINE

Given a user, returns profile metadata and all past tweets

Challenges

Pricing and Limits

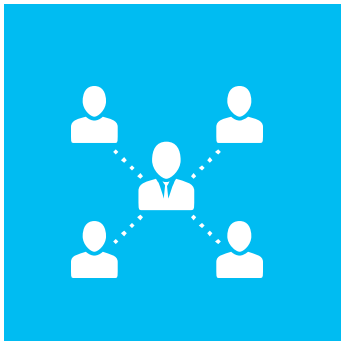
Pricing:

- Firehose access is not easy to get
- Paid access to past tweets

Limits:

- Rate limits
- API limits
- User Timeline to rescue!

Inapplicability of Prior Work



~~Aggregate Estimation over Graphs~~

Issues:

- Broad vs Specific aggregates
- Inefficiency of Random Walks in OSNs

Design Issues

Subgraph Design

Identify the best subgraph for sampling

Sampling Design

Design a topology aware random walk that leverages graph properties

Subgraph Design

Subgraph Design

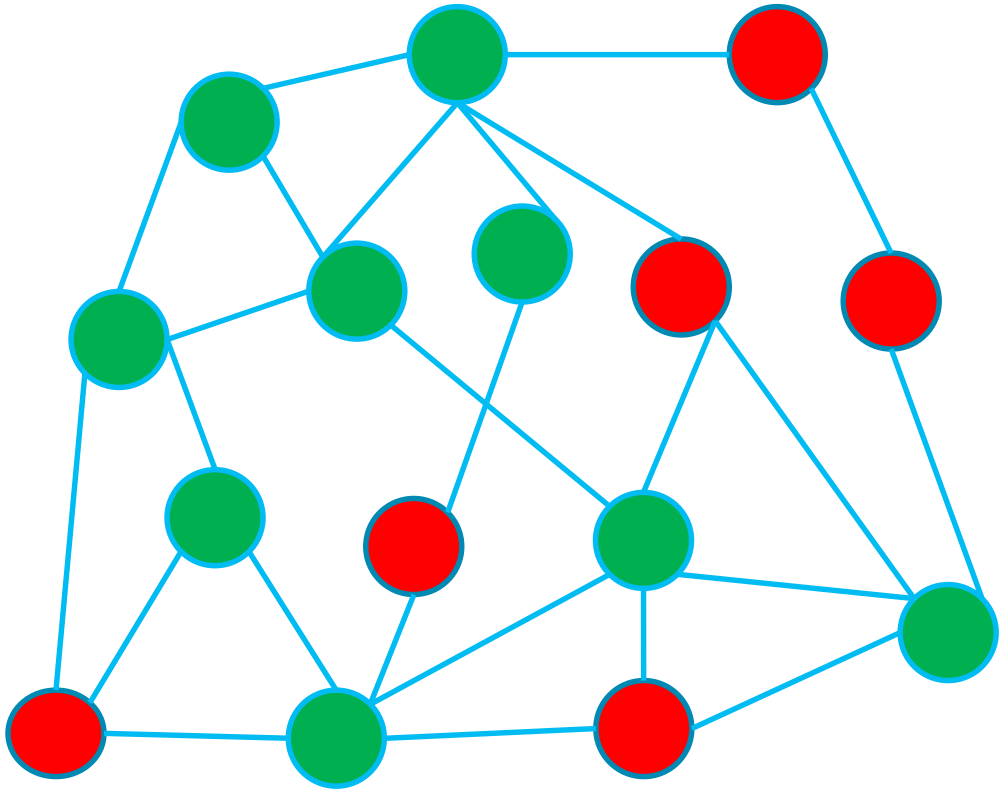
Subgraph Design

Identify the best subgraph for sampling

What is a good subgraph?

- Good Recall
- Topology efficient for sampling

Social Graph



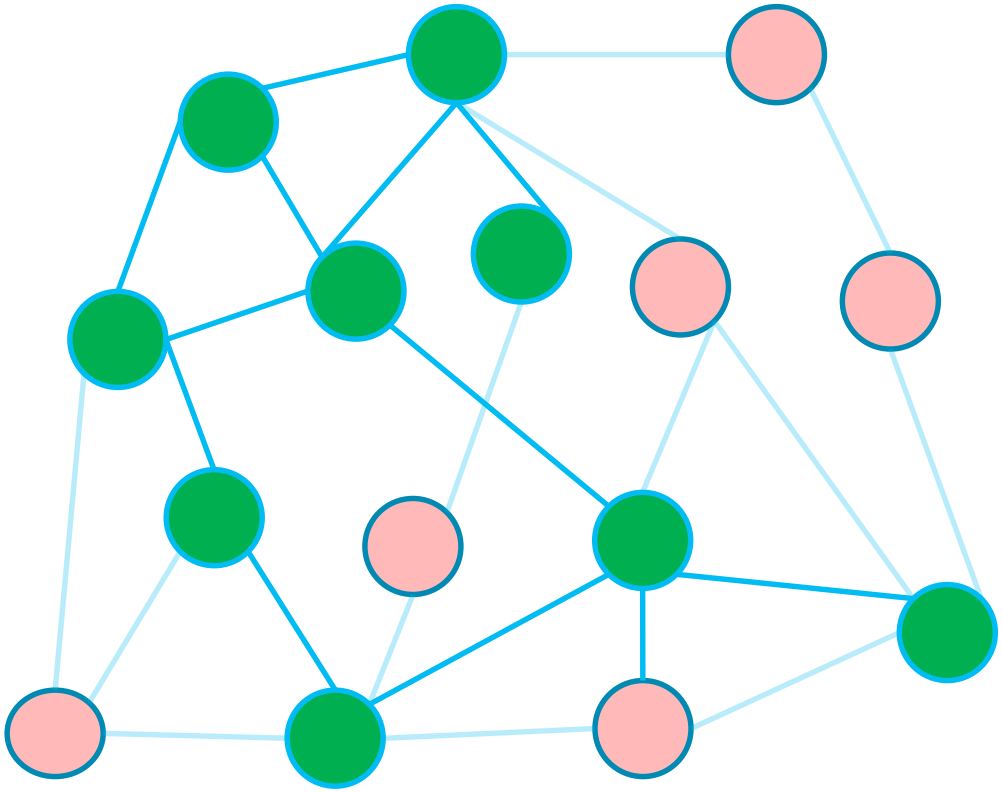
Pros

- Baseline graph
- High recall

Cons

- Very few relevant nodes
- Inefficient for sampling
- High query cost

Term Induced Subgraph



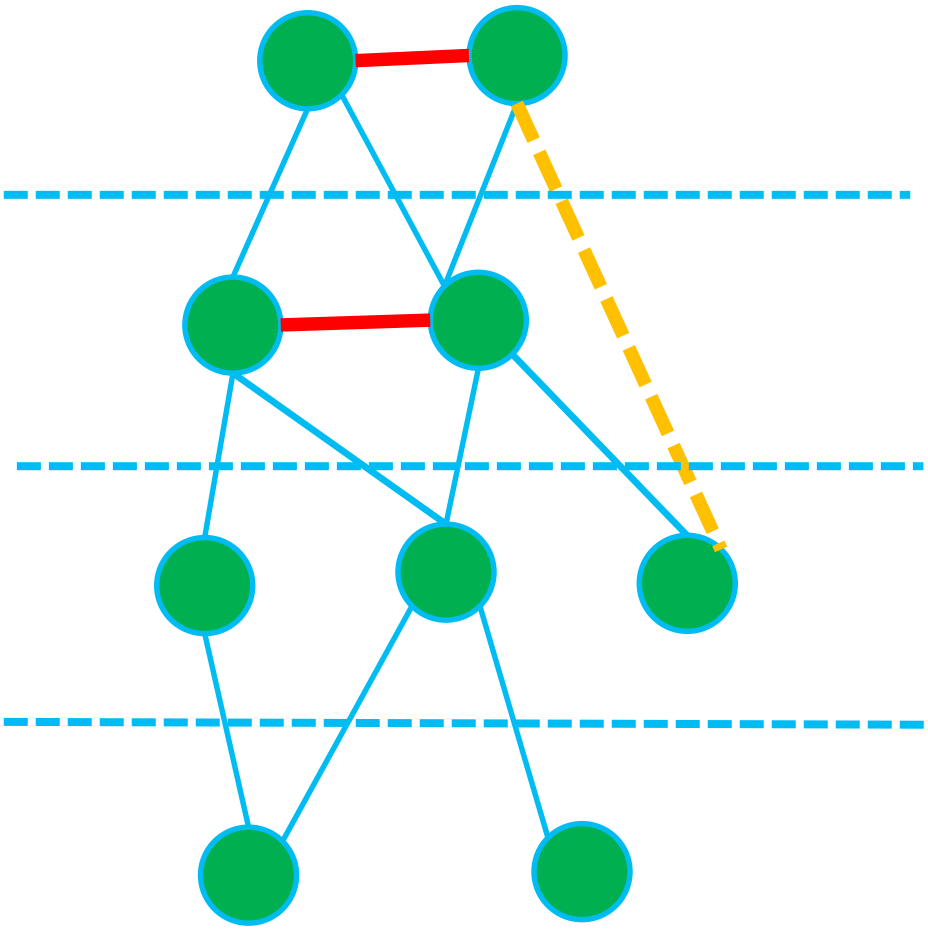
Pros

- Good recall
- Typically well connected

Cons

- Inefficient for sampling
- High query cost

Level by Level Subgraph



- Motivation
  - Communities
  - 90+% retweets within 1 hour
- Intra edges
  - Traps random walks
  - Removing improves conductance
- Inter edges
- Cross edges
- Edge taxonomy and distribution

Removal of Intra Edges

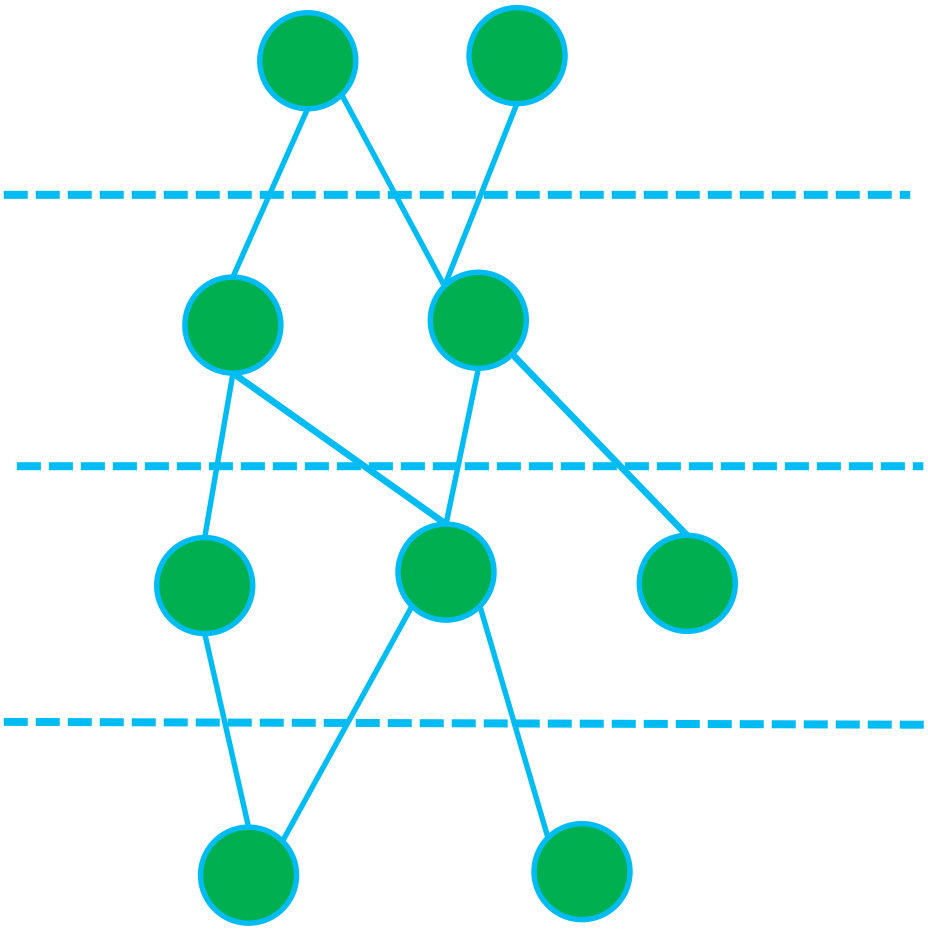
THEOREM 4.1. *The conductance for G is*

$$\varphi(G) = \begin{cases} \frac{h}{(k+d)(h-1)n} & \text{if } d \le \frac{n}{2h},\ k \le \frac{n}{2h} \\ \min\left(\frac{2kh-n}{kh+dn}, \frac{2d}{2d(h-1)+hk}\right) & \\ & \text{if } d \le \frac{n}{2h},\ \frac{n}{2h} < k < \frac{n}{h} \\ \min\left(\frac{2dh-n}{kh+dn}, \frac{2d}{2d(h-1)+hk}\right) & \\ & \text{if } \frac{n}{2h} < d < \frac{n}{h},\ k \le \frac{n}{2h} \\ \min\left((k-\frac{n}{2h})\frac{2dh-n}{kh+dn}, \frac{2d}{2d(h-1)+hk}\right) & \\ & \text{if } \frac{n}{2h} < d < \frac{n}{h},\ \frac{n}{2h} < k < \frac{n}{h} \end{cases} \tag{2}$$

After removing all intra-level edges, the conductance of G' is

$$\varphi(G') = \begin{cases} \frac{h}{nd(h-1)} & \text{if } d \le \frac{n}{2h} \\ \min\left(\frac{2hd-n}{nd}, \frac{1}{h-1}\right) & \text{if } \frac{n}{2h} < d < \frac{n}{h} \end{cases} \tag{3}$$

Level by Level Subgraph



Pros

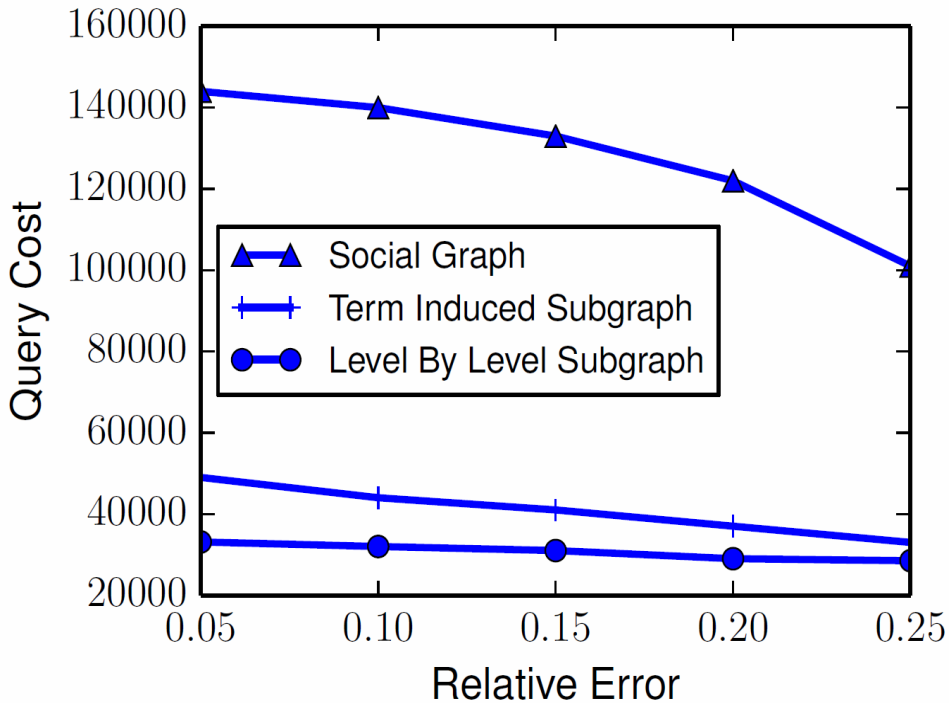
- Recall comparable to term induced subgraph
- Reasonably connected
- Better efficiency for sampling
  - Sparser graph
- Lower query cost

Algorithm So Far

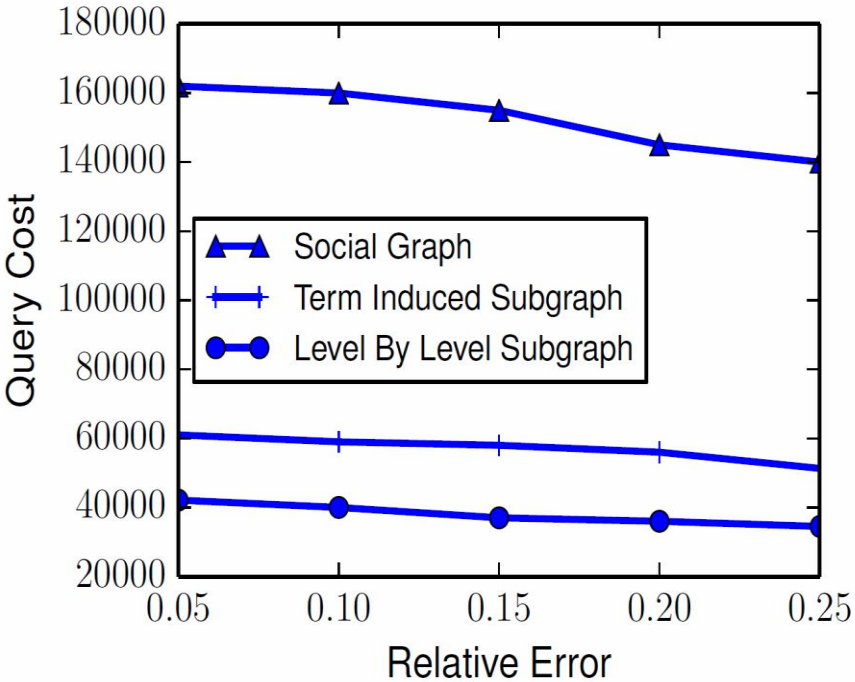
Algorithm MA-SRW

1. Get seed users
2. Repeat until budget is not exhausted
   - Retrieve samples via SRW
   - Pick only edges that belong to another level
3. Perform aggregate estimation over collected samples

MA-SRW Performance



- AVG(followers)
- Keyword=privacy
- Year=2013



- COUNT(users)
- Keyword=privacy
- Year=2013

Sampling Design

Sampling Design

Sampling Design

Design a topology aware random walk that leverages properties of Level-by-Level graph

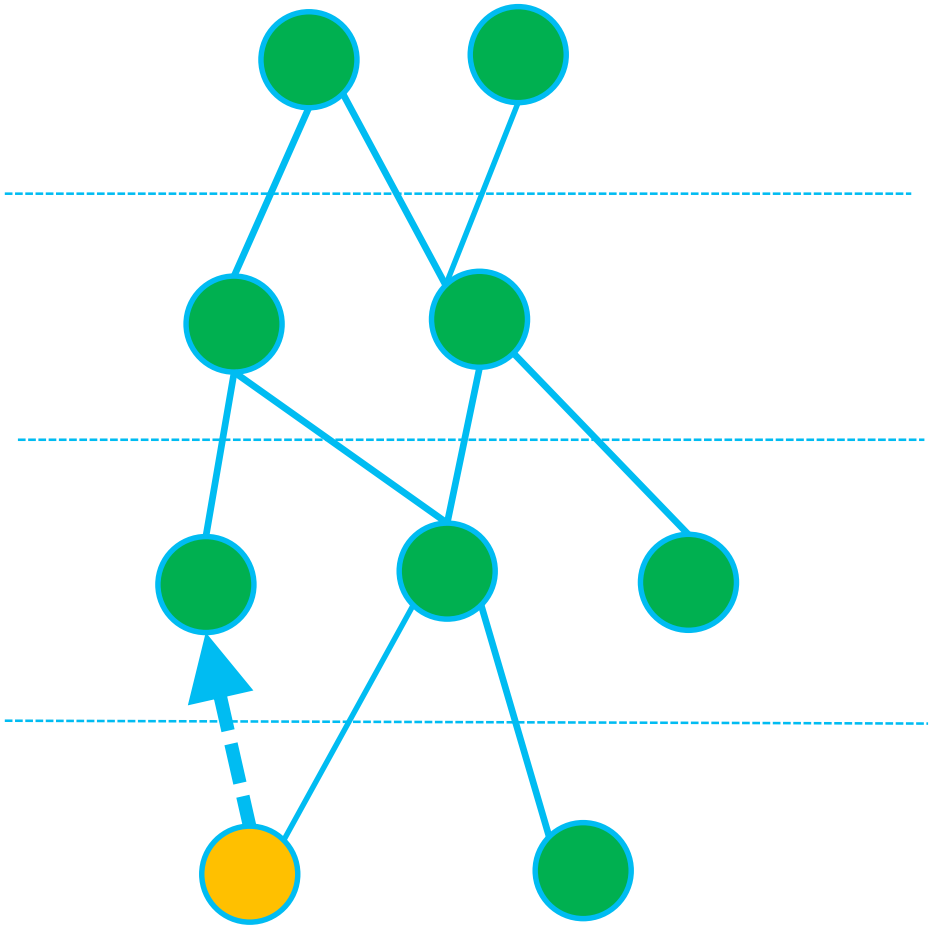
Deficiencies of Random Walks

1. High query cost for SUM/COUNT aggregates
   - Mark and recapture might require $\Omega(\sqrt{n})$ samples for one collision

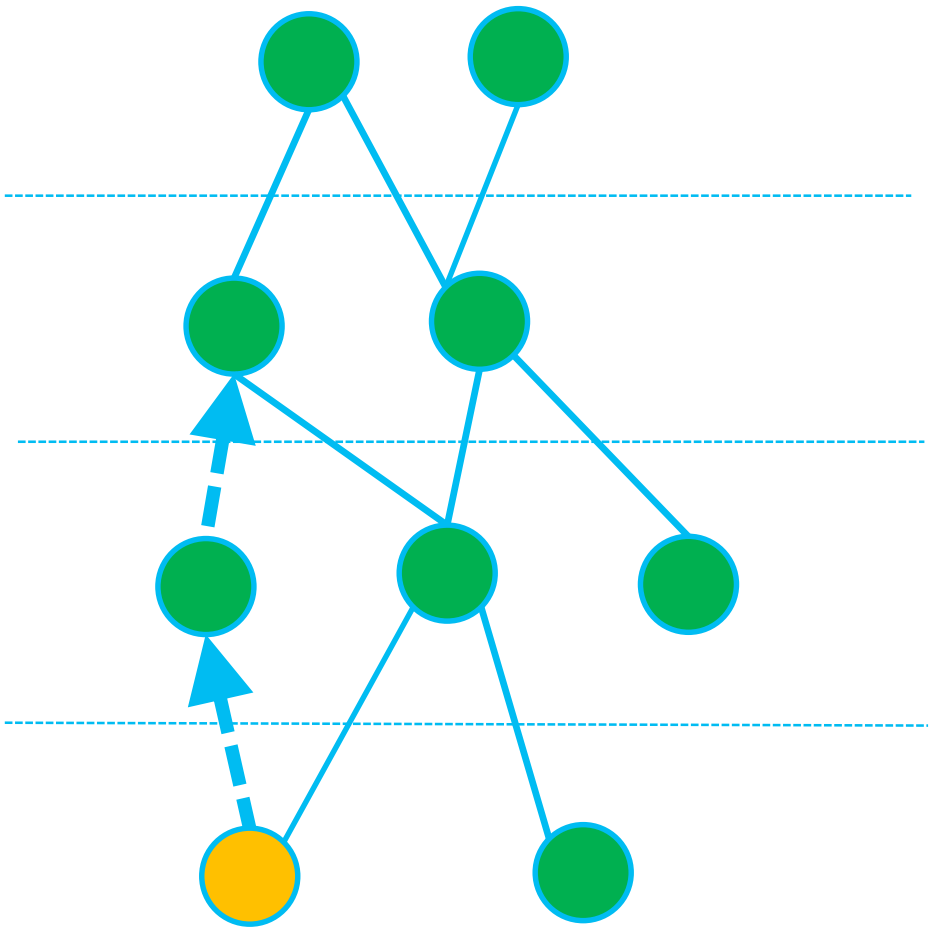
2. Large burn-in despite asymptotic unbiasedness
   - Especially bad for online social networks
   - Oblivious to topology

3. Can we compute the probabilities proactively?
   - instead of waiting till convergence to stationary distribution?

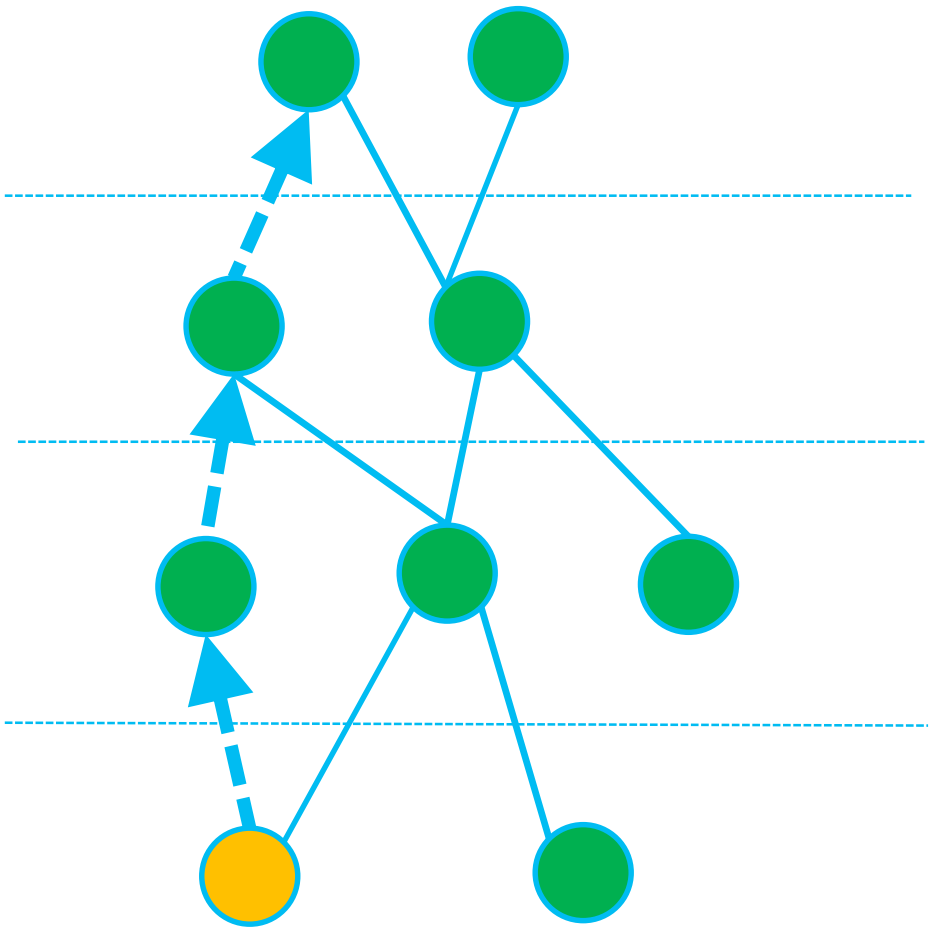
Level by Level Random Walk



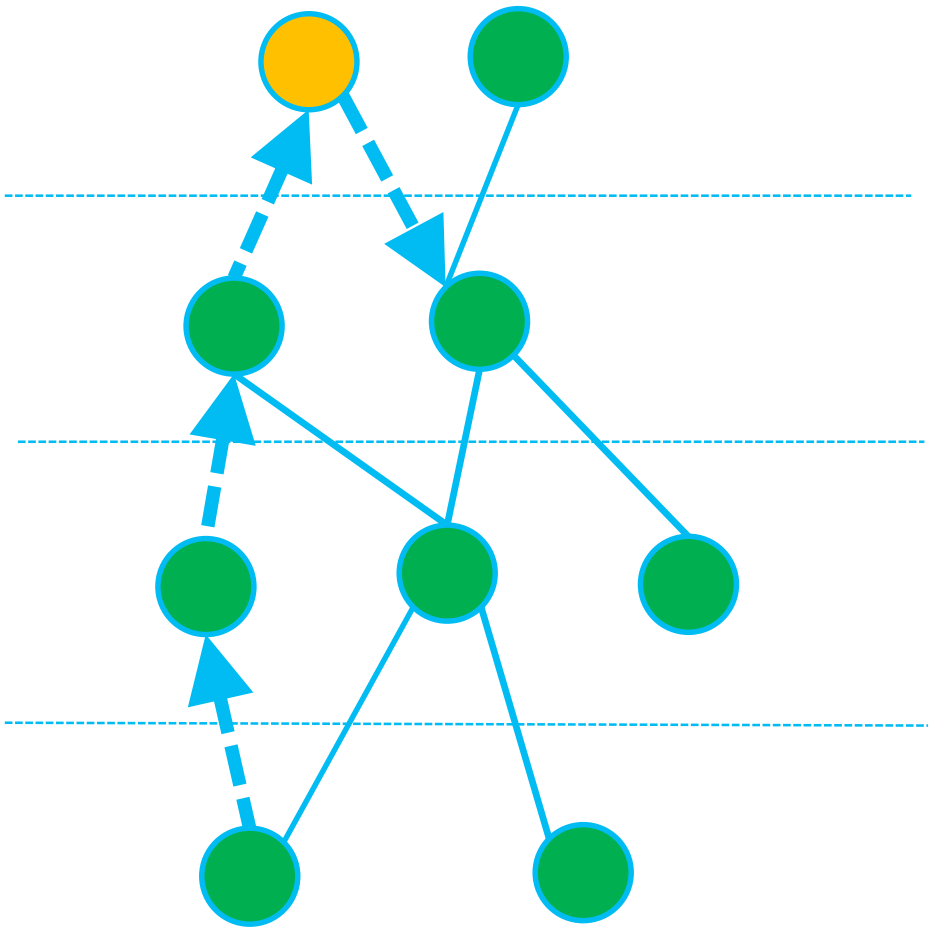
Level by Level Random Walk



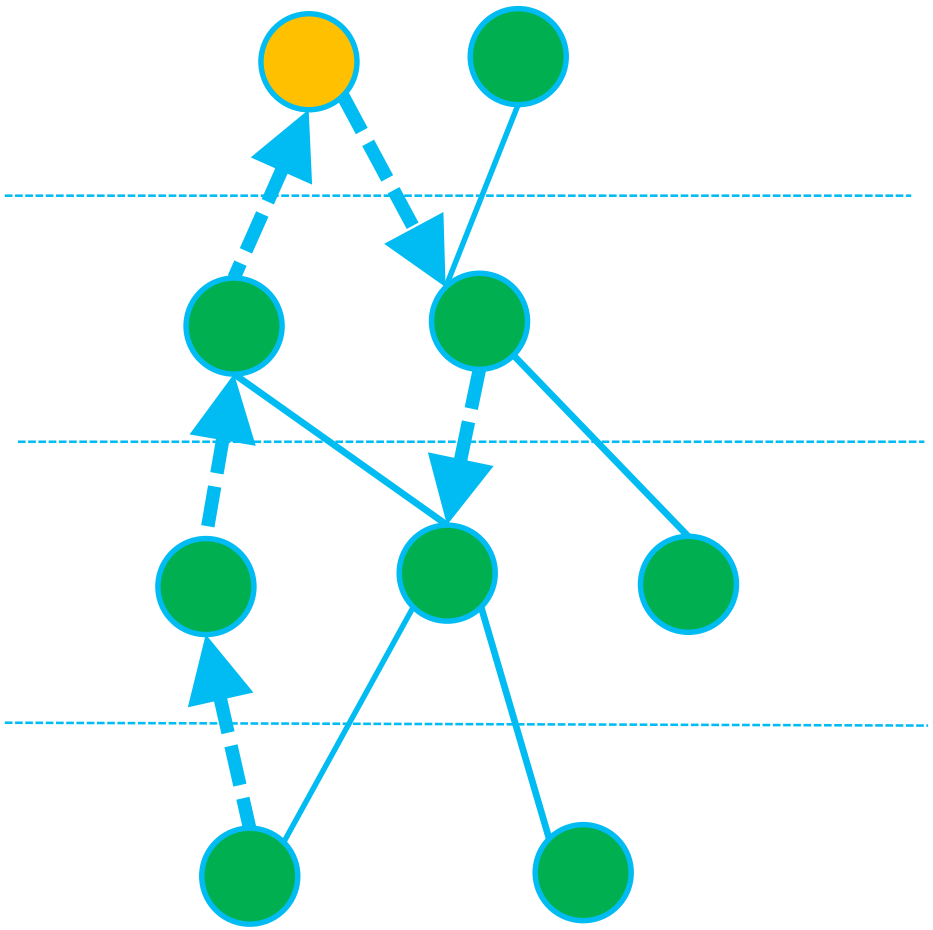
Level by Level Random Walk



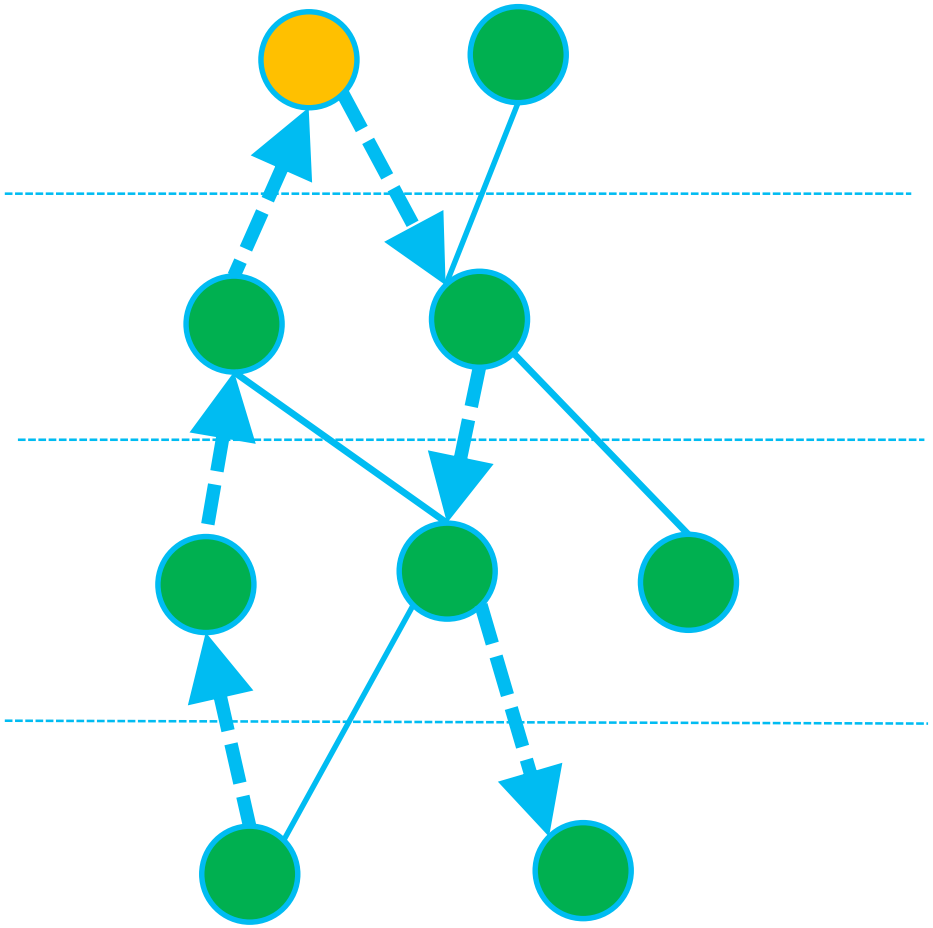
Level by Level Random Walk



Level by Level Random Walk



Level by Level Random Walk



How to estimate probabilities?

$$\acute{p}(u) = \sum_{v \in \Delta(u)} \frac{\acute{p}(v)}{|\nabla(v)|}, \qquad \grave{p}(u) = \sum_{v \in \nabla(u)} \frac{\grave{p}(v)}{|\Delta(v)|},$$

Unbiased Estimator!

$$\omega(\acute{p}(u)) = \frac{|\Delta(u)| \cdot \acute{p}(v)}{|\nabla(v)|}$$

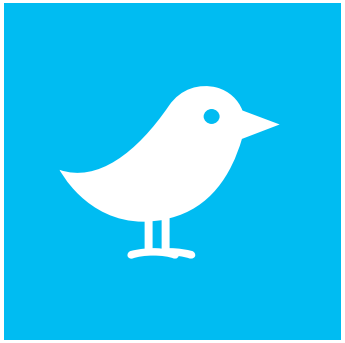
Final Algorithm

Algorithm MA-TARW

1. Get seed users
2. Repeat until budget is not exhausted
   - Perform bottom-top-bottom random walk R_i
   - Estimate sampling probabilities for all nodes in R_i
   - Estimate aggregate using all nodes in R_i via Hansen-Hurwitz estimator
3. Return average of all previous estimates

Experiments

Datasets



Twitter

Google+

Tumblr

Experiment Details

Keywords

- Privacy, New York, Boston
- Cities, Celebrities, Companies, popular and niche hashtags

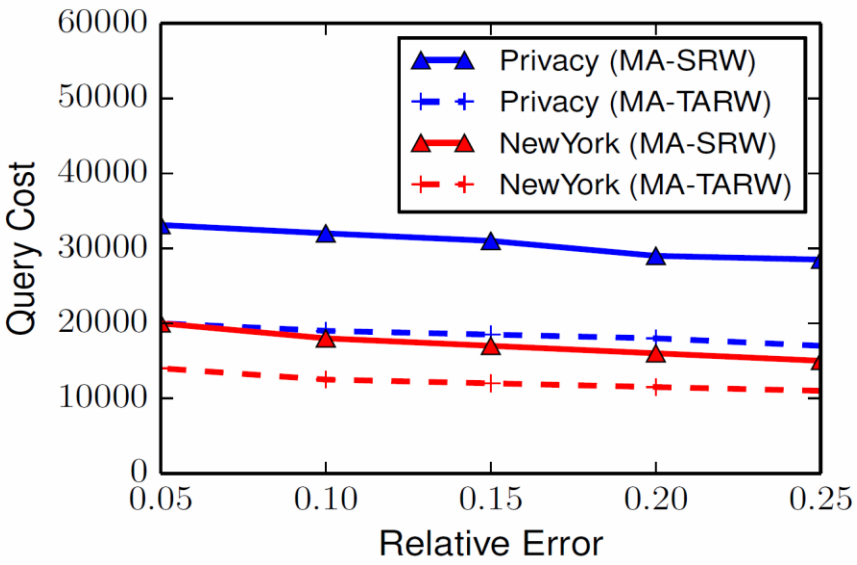
Aggregates

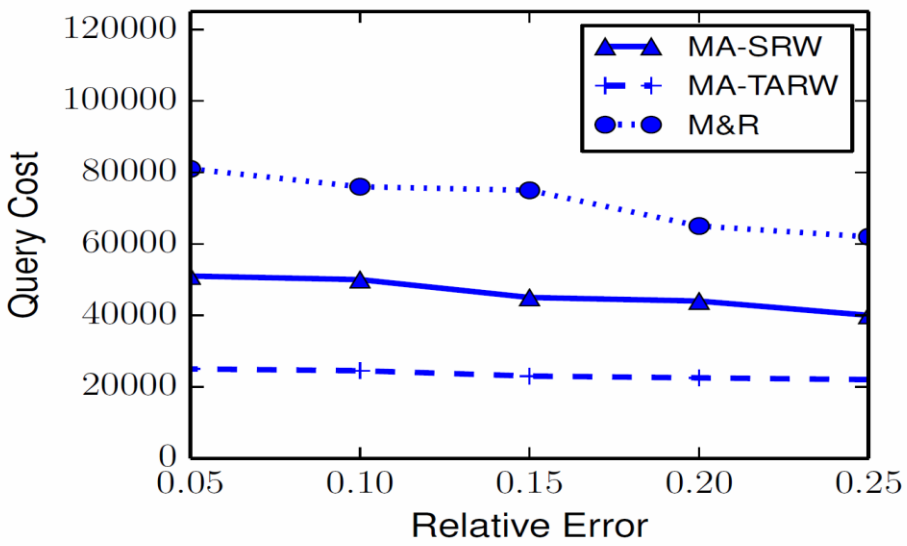
- AVG(followers)
- AVG(display name)
- COUNT(male)

Time

- Restrict tweets from 2013

Experimental Results

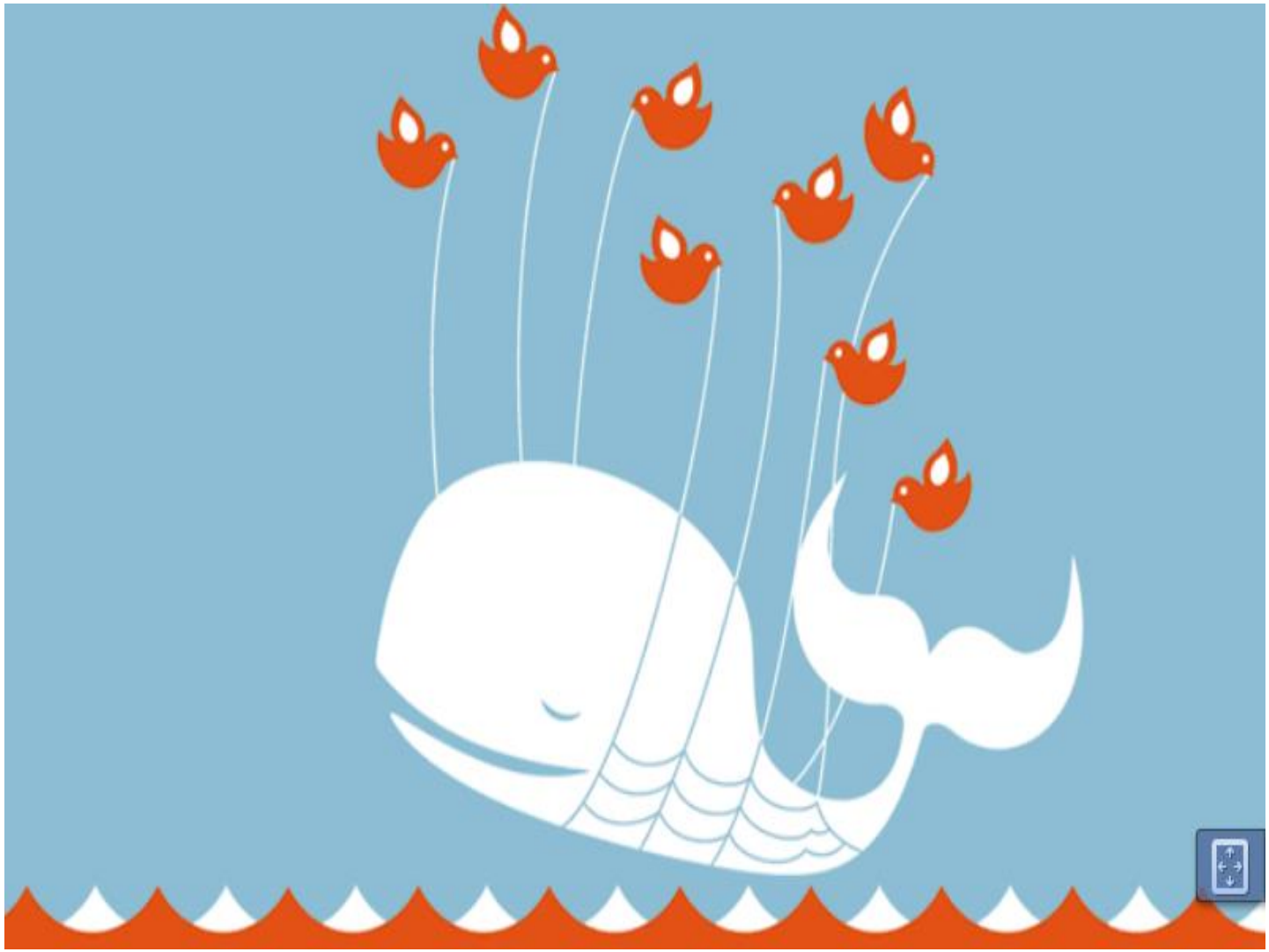




- AVG(followers)
- Keyword=privacy
- Year=2013

- COUNT(users)
- Keyword=privacy
- Year=2013

Summary

- Aggregate Estimation over Microblogs
- Subgraph Design
- Sampling Design

Questions?