Influence Maximization with Fairness at Scale

Michael Golas, Emily Robles, Abhi Sharma, Justin Wong

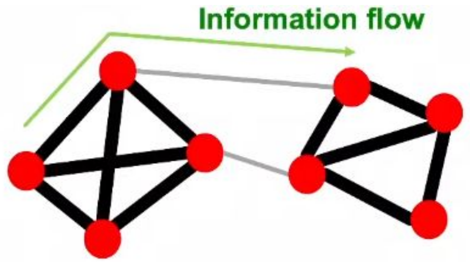
W210 | Presentation 1 | Week 5

Berkeley
UNIVERSITY OF CALIFORNIA

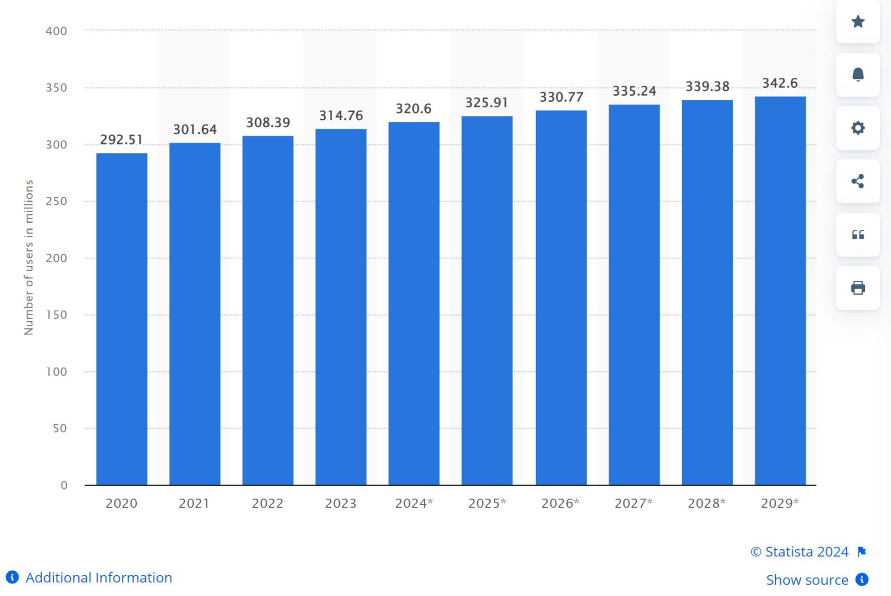
Definitions and Motivation

Social Networks (Graphs)

- **Omnipresent** in our lives, information spread is massive but **unfair**
  - Can create **information asymmetric (dis)advantage**
    - Stock news, vaccine availability, targeted violence
- Long range edges allow for gathering of information from different parts of the network
- Social connections can influence individual characteristics of a person (homophily - **birds of the same feather** flock together)

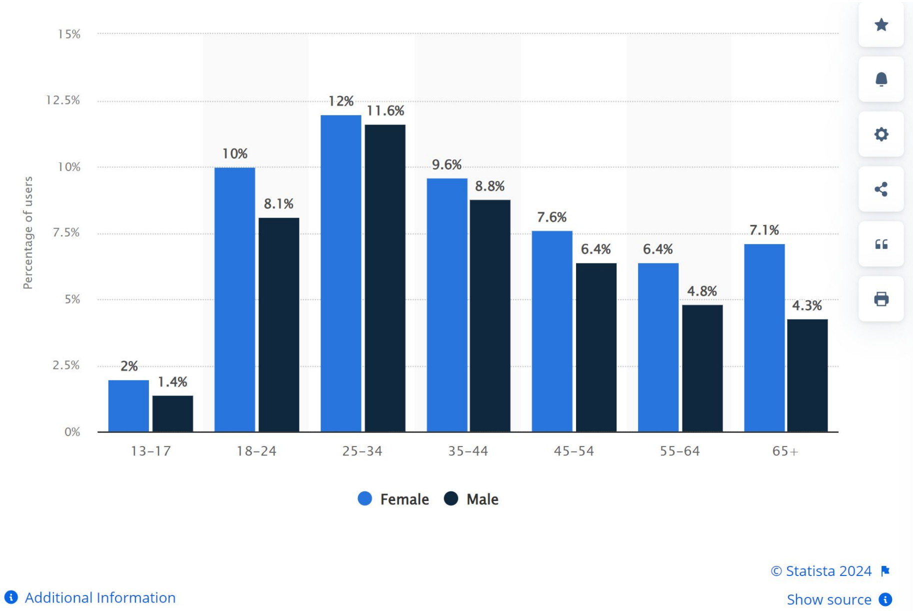


Berkeley
UNIVERSITY OF CALIFORNIA



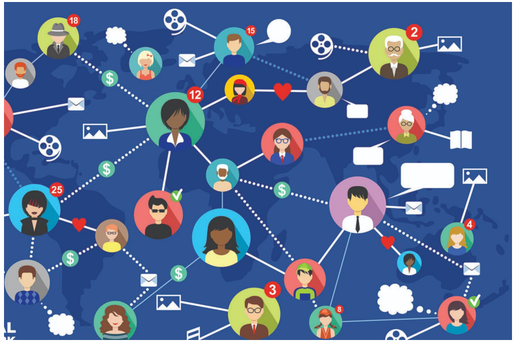
Users of Social Media (USA)

FB Users by Age Group and Gender (USA Aug 23)



Influence Maximization

- Class of algorithms that aim to **maximize information spread in a graph** under some **budget constraints**
- Find **K most influential (seed) nodes** from which diffusion of a **specific message** should **start**
- **Examples**: Effective marketing, targeted social media
- **Goal:** Not only maximize info, do it preserving **"fairness"**



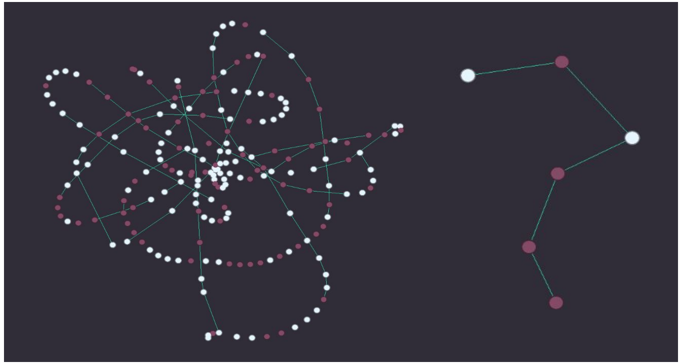
Influencers

- Set of **K nodes** (people) in the graph
- Have individual characteristics **(dimensions)**
  - Age, Gender, Location, Race, Number of Followers
  - Influencer characteristics **may differ** from network's characteristics
- Have high "trust" (voice) in the network - potential to **"maximize influence"**
  - Their tweet can cause a **diffusion cascade**

Diffusion Cascade

- **Diffusion:** the process of **spreading** a message in a network
  - Expected spread: sum of probabilities
- **Cascade:**
  - A collection of ***(source, target, timestamp)*** for a **specific message** *(post)*
  - A **replay** of the network spread of a message
  - **Chain reaction:** largest is 50k retweets (Weibo)
  - Influencers are **earliest timestamp** in cascade

Large vs. small influence



“Fairness” Constraint

- **Group:** graph is divided into groups of people with similar **characteristics (age, gender)**
- **During diffusion** of message, ensure **characteristics are fairly affected**
- **Diversity is preserved:** Every group *(age, gender)* receiving influence is commensurate to what it would generate on its own
    - Does the influenced graph “look the same” from ***influencer spread*** vs ***internal spread***
- Maximizing influence for one group can **reduce it for another**
    - Subset of population (old males) was informed on vaccine availability

"Fairness" Constraint

- **Equality:** each group gets the same number of seeds (influencers)
- **Maxmin:** *minimize gap* of information spread *between groups*
    - **70% of old males** and **69% of young females** influenced with *same message*
    - **Proportions are preserved** between groups
- **Equity:** any person's probability of being influenced is (almost) the same, regardless of group
    - Vaccine news will equally reach old man and young woman
    - **Demographic parity**

Problem Formulation

We want to **maximize the information spread** of a social message

- Example: covid vaccine availability and eligibility of administration

Using **Weibo's influencer set** subject to **fairness constraints** of **equity and diversity**

Or More Generally...

Given a network with **N nodes** and given a “spreading” or **propagation process** on that network, choose a **“seed set” S of size k < n** to **maximize the number of nodes** in the network that are ultimately influenced under the **“fairness” constraint**.

Berkeley
UNIVERSITY OF CALIFORNIA

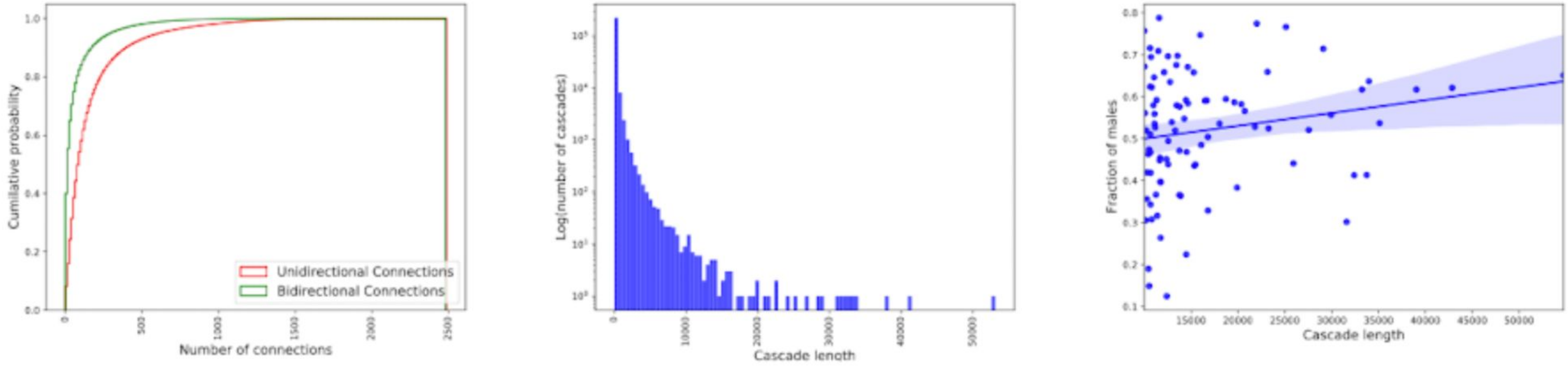
Dataset: Sina Weibo

Dataset Description

- Chinese Social Media network (like Twitter)
  - 1.8M users, 308M relationships (in dataset - 2012)
  - Founded 2009, **Weibo** is chinese for **“microblogging”**
  - One of the largest social networks (**252M DAU**), $30B market cap (2018)

Users	Follow Relationships	Original Microblogs	Retweets
1,776,950	308,489,739	300,000	23,755,810

Dataset: Sina Weibo



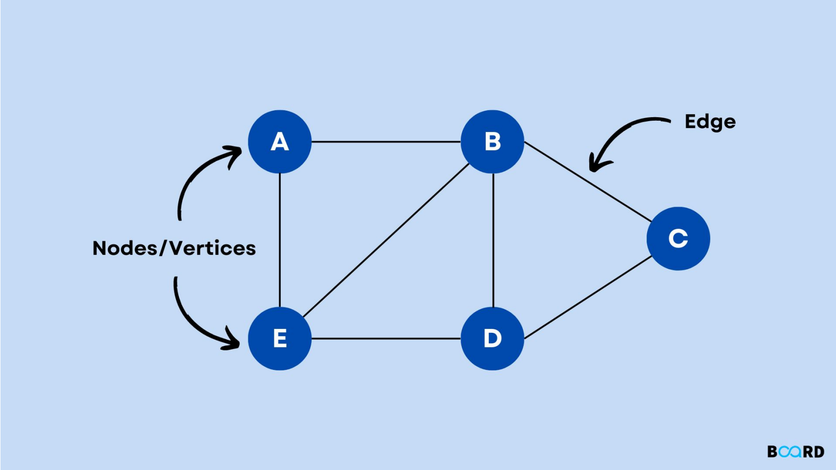
(a) (b) (c)

Figure 4: Statistics of the Weibo dataset: (a) distribution of social connectivity, (b) distribution of cascades, (c) gender distribution in long cascades.

Components of an IM Algorithm

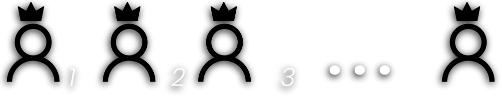
1. **Network Graph**
2. Budget *(Constraint)*
3. Influence Model
4. Optimization Framework

   (Seed Selection)



1. Network Graph
2. **Budget *(Constraint)***
3. Influence Model
4. Optimization Framework

   (Seed Selection)

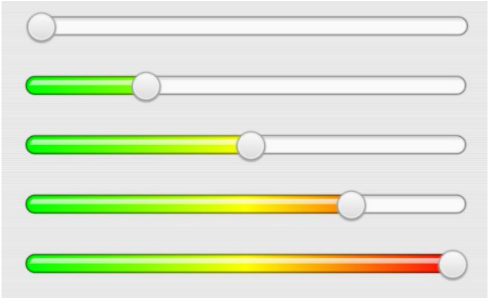


"K" Influencers

Fairness of characteristic "s"
CV of "influenced ratios"

$$CV = \sigma/\mu,$$

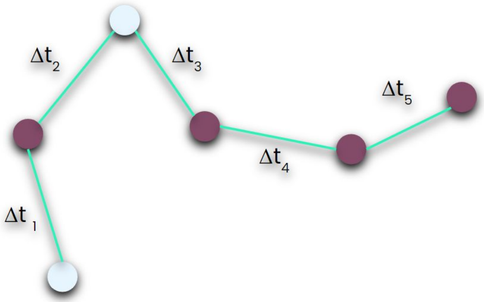
$$f_s = \frac{2}{1 + \exp(CV)}.$$

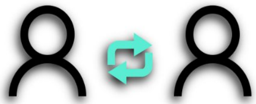


Fairness Threshold

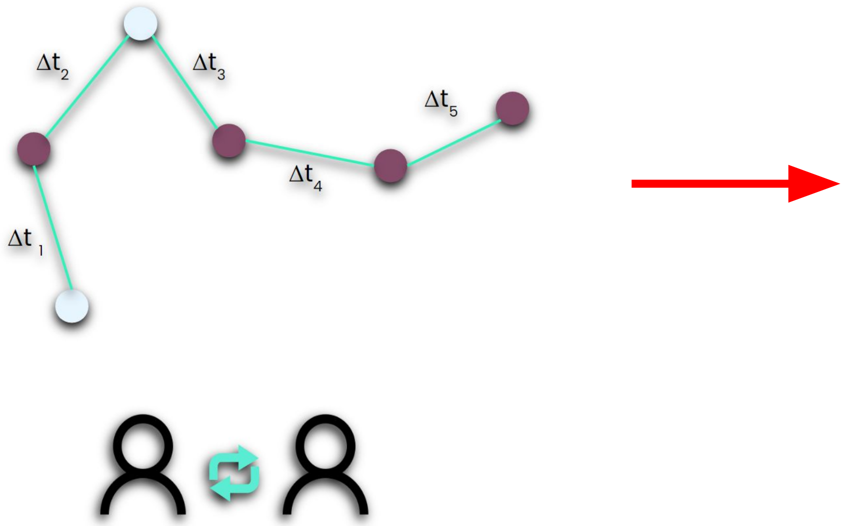
1. Network Graph
2. Budget *(Constraint)*
3. **Influence Model**
4. Optimization Framework

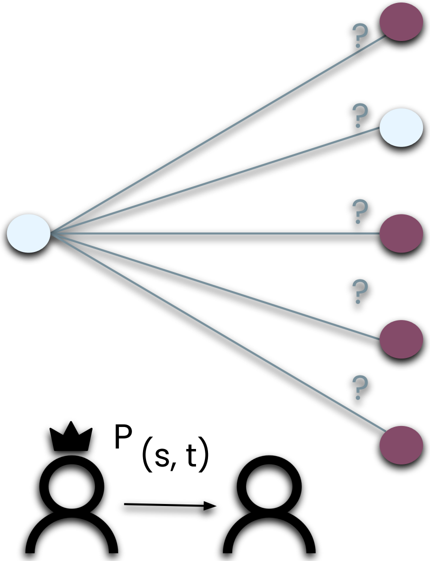
   (Seed Selection)





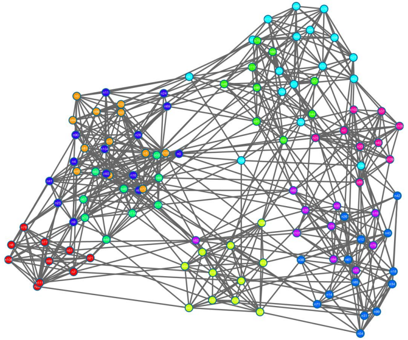
- Use diffusion cascade data observed
- **Eliminate distance** between nodes
  - Interested in **influenced** indicator **(Y/N)**
- Convert to **bi-partite graph**
  - Influencer to affected users
  - **Key differentiation**





P (s, t) Diffusion Probability

- **Diffusion Probability P(s, t):** Probability that **node "t"** will be *found in cascade* started by **node "s"**
  - So *t* is influenced by *s*
- Why convert to **bi-partite graph**?
  - **Assume connectivity** between *s* and *t* is always there ***(direct or indirect)*** - so *eliminate exploration* of possible connections between nodes
    - If probability is 0, no connection found
    - Reduce computational complexity!
      - **Scale to millions of nodes!**



Berkeley UNIVERSITY OF CALIFORNIA

Next Time...

- Fairness-based Participant Sampling (FPS)
- Fairness As Context (FAC)

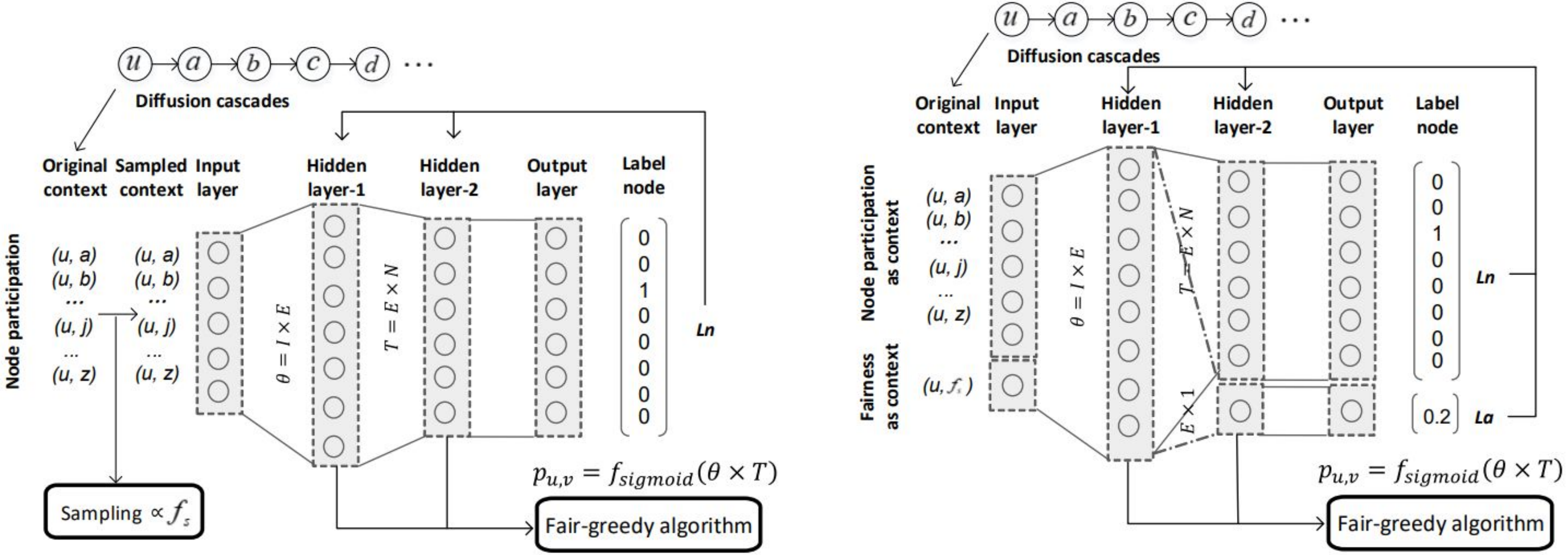
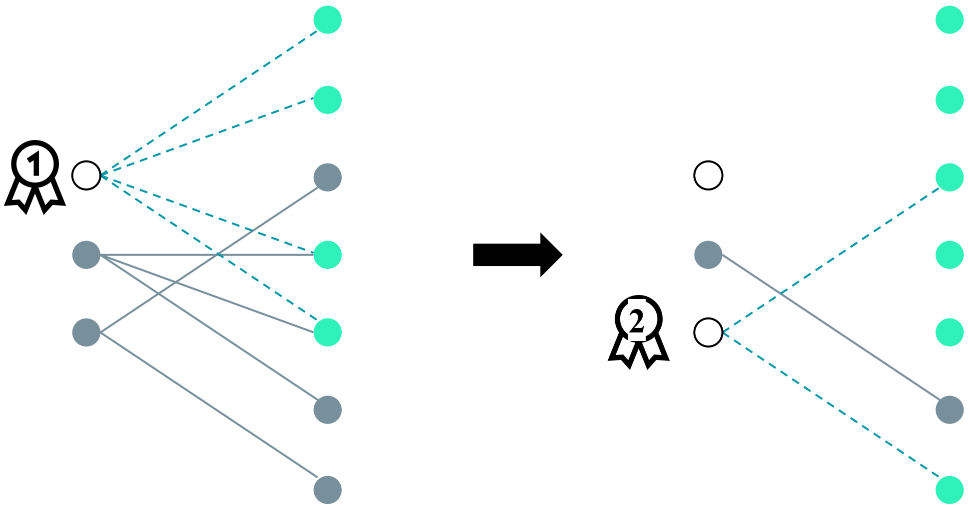


Figure 1: Illustration of our algorithmic solutions: FPS (left) and FAC (right).

1. Network Graph
2. Budget *(Constraint)*
3. Influence Model
4. **Optimization Framework *(Seed Selection)***

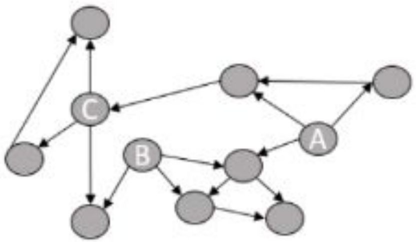
Modified Cost Effective Lazy Forward

Rank top seeds that maximize influence spread, remove, & repeat

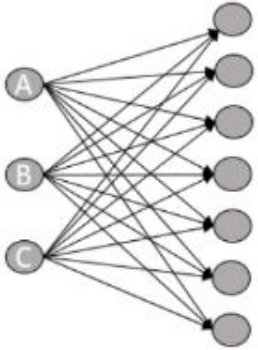


Summary

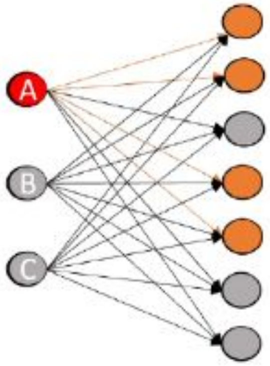
Create network from social media retweets



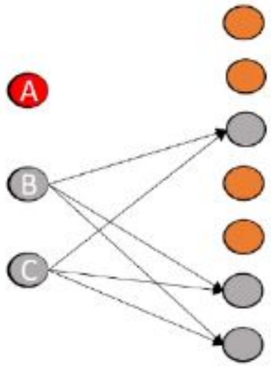
Model network as bipartite graph (influencers vs. other users)



Derive cascade probabilities from neural network



Select top influencers that maximize spread of information, remove, & repeat



Road Ahead …

Challenges Ahead

- Replication of code results
  - Validate the solutions by comparison with previous results
  - Confirm generation of results within epsilon
  - Fairness gains do not come at the expense of influence
  - Algorithm complexity $O(k|I||V|\log|V|)$ on E5-2620 with 188 GB RAM and Tesla k40m GPU
    - require relatively high degree of data caching
    - data proximity important
  - Digg ≈ 1% of Weibo - use for testing
    - FPS performs relatively worse in Weibo vs. Digg

Proposal and Deliverables

- Replicate paper results to establish baseline and begin experimenting with refinements to the thesis, and exploration of results
  - Improve actual (non-asymptotic) runtime, perform larger number of repeatable experiments to test various hypotheses related to fairness

- Make incremental changes to neural network architecture
  - FPS and FAC
  - Skip connections
- Optimization: multi-objective optimization
  - Aggregative training vs. concatenation of node embeddings
- Explore non-categorical variables
  - Age
  - Number of followers
  - Average post length

Team Composition

Name	Role
Emily	Project Management; FAC/FPS Research
Abhi	Neural Network Exploration (FAC, FPS)
Michael	Coding; EDA, Data Processing
Justin	Coding; Evaluations

Berkeley
UNIVERSITY OF CALIFORNIA

Questions?

Appendix

Appendix

Question: Computational complexity savings on re-formulation of graph as bi-partite?

- This actually results in a large computational savings
- C = cascade length, N = total nodes in system
- Building the context between two nodes on the retweet network would have **O(C*N(N-1)/2)** complexity versus **O(C*N)** complexity for the bipartite representation in creating the training examples
- **The previous process requires the creation of the propagation network, meaning going through every node** in the cascade and iterating over the subsequent nodes to search for a directed edge in the network. This has a complexity of $O(c(\bar{n}(\bar{n}-1)/2))$, where c is the number of cascades and $\bar{n}$ is the average cascade size. Given that the average size of a cascade can surpass 60 nodes, it is a very time consuming for a scalable IM algorithm.
- The node- context creation has a complexity of O(cn), which is linear to the cascade's size and does not require searching in the underlying network.

Berkeley
UNIVERSITY OF CALIFORNIA

Appendix

Question: how to calculate fairness of a single user?

1. **These are aggregate statistics on the right**
2. **So for a single user "u", we take equation 5 and calculate CV using data such that "u" is the initiator**

$$CV = \sigma/\mu, \tag{2}$$

where σ, the standard deviation of the influenced ratios, is:

$$\sigma = \sqrt{\frac{\sum_{i \in C_s} \left(\frac{|\Omega_i|}{|V_i|} - \mu \right)^2}{|C_s|}}, \tag{3}$$

and μ denotes the average of influenced ratios:

$$\mu = \frac{1}{|C_s|} \sum_{i \in C_s} \frac{|\Omega_i|}{|V_i|}. \tag{4}$$

With the relative dispersion of influenced users in the groups induced by s to capture unfairness, the fairness score can then be scaled by a *sigmoid* function and bounded between 0 and 1, by[1]

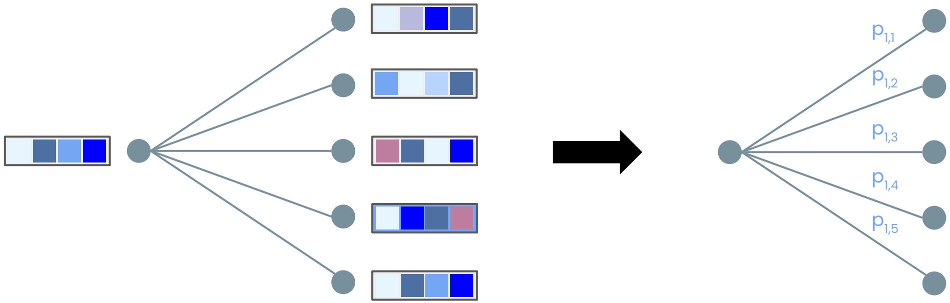
$$f_s = \frac{2}{1 + \exp(CV)}. \tag{5}$$

Berkeley
UNIVERSITY OF CALIFORNIA

Components of an influence maximization algorithm

1. Network graph
2. Budget
3. **Influence model**
4. Optimization framework

Representation learning with cascade context to infer diffusion probabilities between seed and target users



Cascade context

Number of participating users and its fairness score
[seed user, target user, # usrs, fair score]

FPS (Fairness-based Participant Sampling)

- Implements a fairness-based penalty
- **Input**: one-hot encoding indicating whether each node was influenced by given node *u*
- **Output**: vector with probabilities that a given node will influence the other nodes

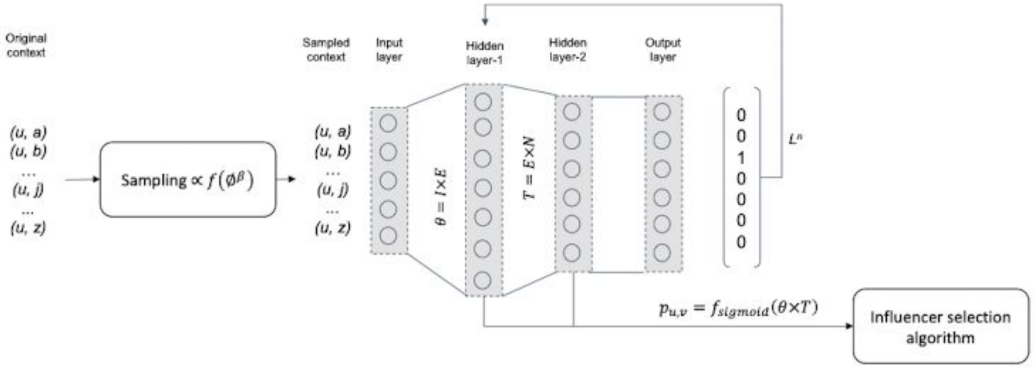


Figure 1: Schematic representation of the FPS model.

FAC (Fairness as Context)

- Implements two separate neural networks (with shared hidden layer)
- **Input for top NN**: one-hot encoded node participation contexts
- **Input for bottom NN**: weighted average fairness of a given influencer
- **Output**: vector with probabilities that a given node will influence the other nodes

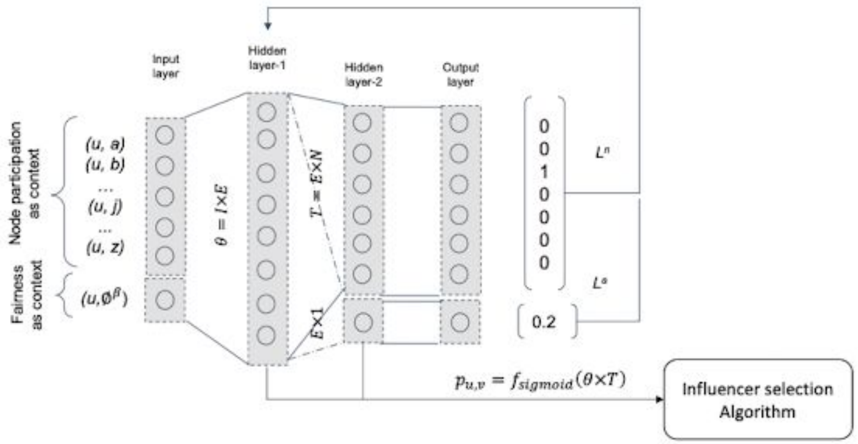


Figure 2: Schematic representation of the FAC model.

Berkeley
UNIVERSITY OF CALIFORNIA