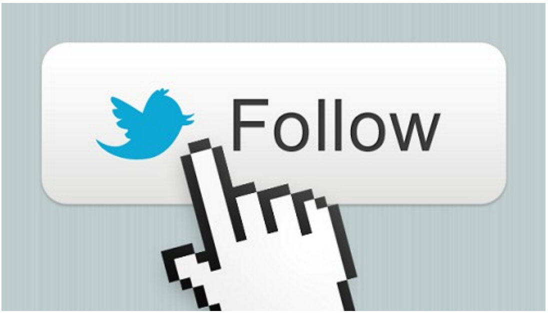
Controlling Information Flow in Online Social Networks

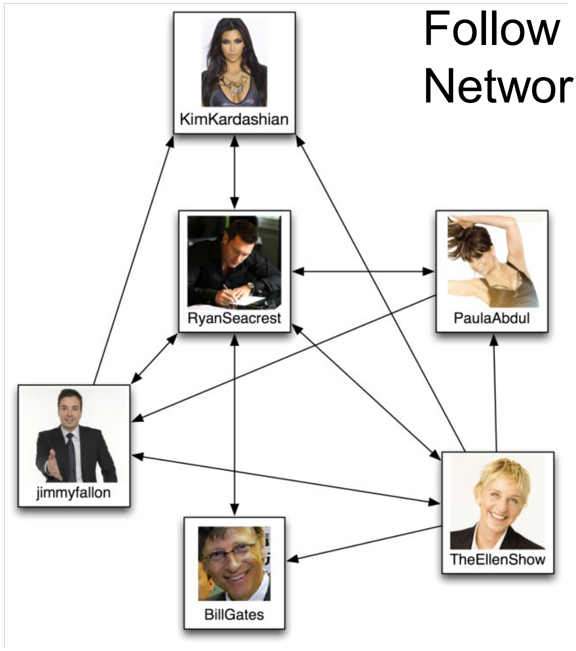
Soumajit Pramanik & Bivas Mitra

Department of Computer Science & Engineering,
Indian Institute of Technology, Kharagpur, India

Introduction

- Follow Links in Twitter
  - A user can be followed by any number of users
  - All tweets by the user are shown in the timeline of her followers.

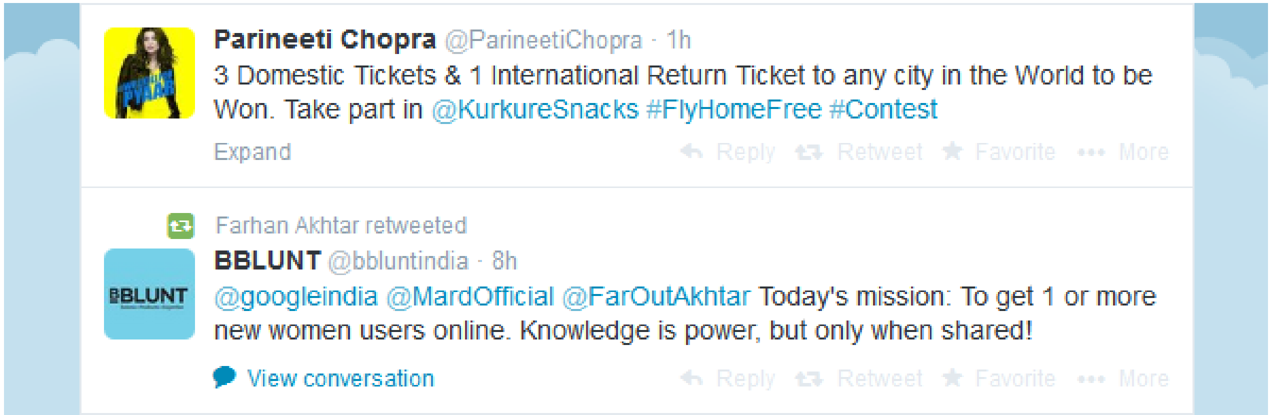




Follow Network

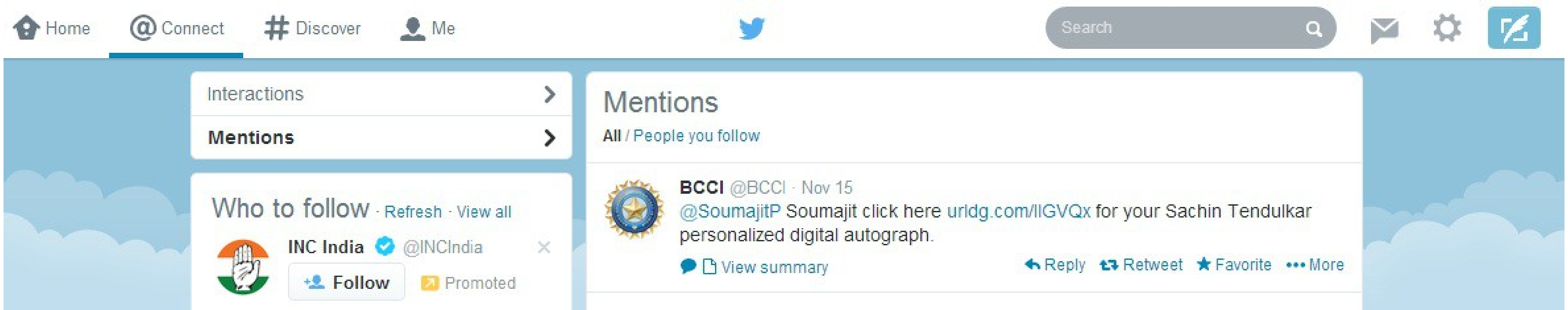
Introduction (Continued)

- Mention Links in Twitter
  - A mention is any Twitter update that contains "@username" anywhere in the body of the Tweet.
  - Non-Followers can also be mentioned



Parineeti Chopra @ParineetiChopra · 1h
3 Domestic Tickets & 1 International Return Ticket to any city in the World to be Won. Take part in @KurkureSnacks #FlyHomeFree #Contest

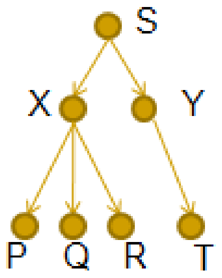
Farhan Akhtar retweeted
BBLUNT @bbluntindia · 8h
@googleindia @MardOfficial @FarOutAkhtar Today's mission: To get 1 or more new women users online. Knowledge is power, but only when shared!



BCCI @BCCI · Nov 15
@SoumajitP Soumajit click here urldg.com/lIGVQx for your Sachin Tendulkar personalized digital autograph.

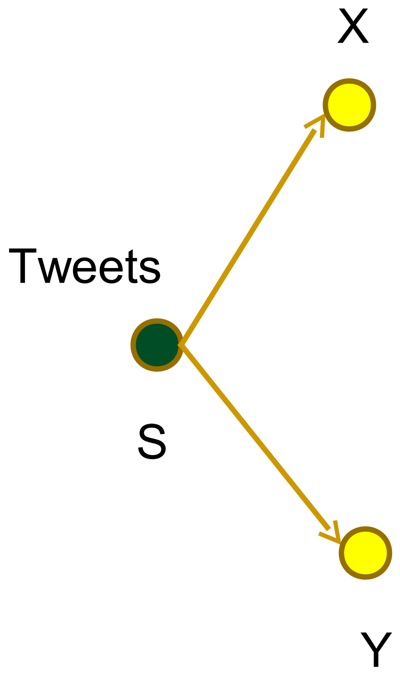
Introduction (Continued)

Information Diffusion via Follow Links



Follower Network

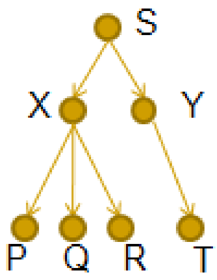
S has two followers X & Y



Hashtag is a unit of information in Twitter

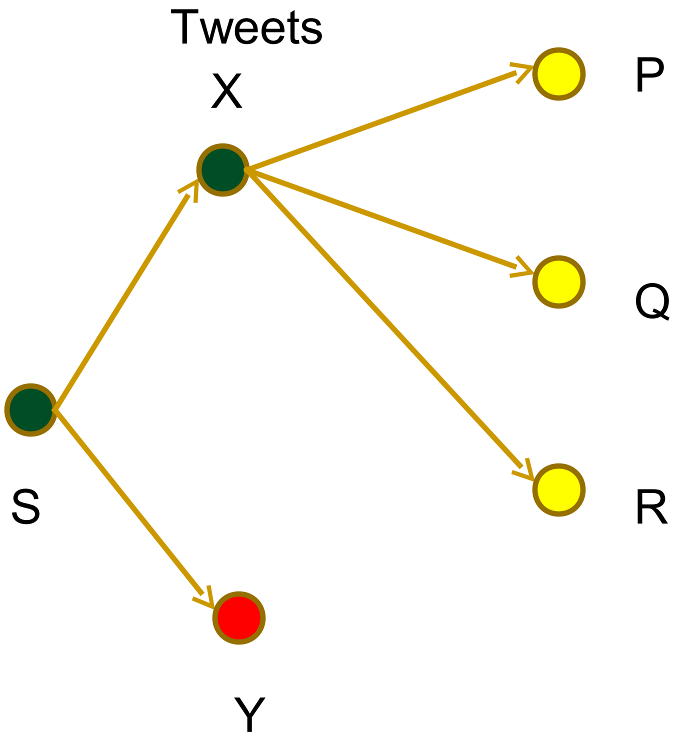
Introduction (Continued)

Information Diffusion via Follow Links



Follower Network

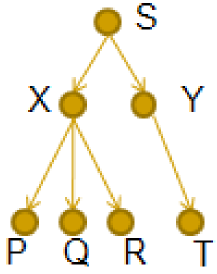
X has 3 followers- P, Q & R



X decides to re-tweet
Y decides not to re-tweet

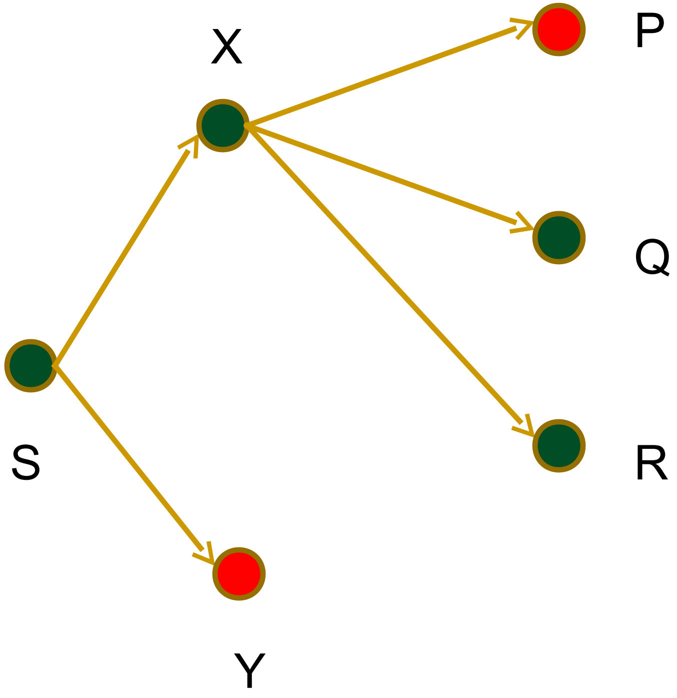
Introduction (Continued)

Information Diffusion via Follow Links



Follower Network

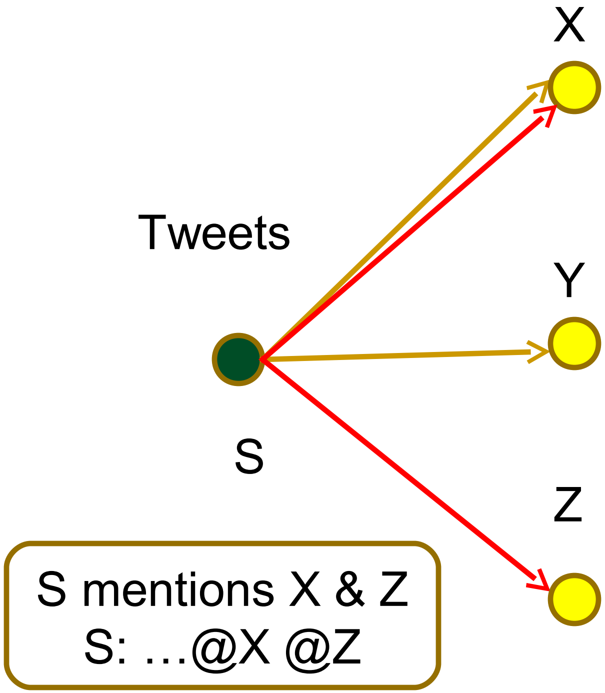
P decides not to re-tweet
Q & R decide to re-tweet



And in this way the information propagates….

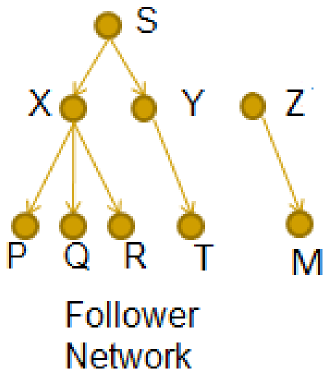
Introduction (Continued)

Information Diffusion via Follow & Mention Links



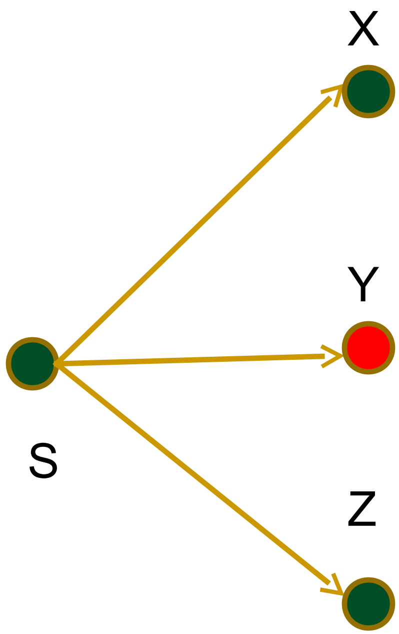
S mentions X & Z
S: …@X @Z

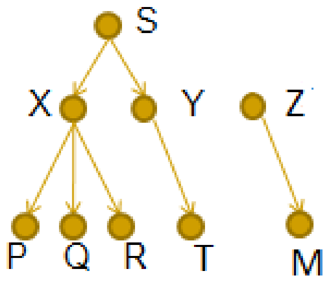
S has two followers X & Y



Introduction (Continued)

Information Diffusion via Follow & Mention Links





Follower Network

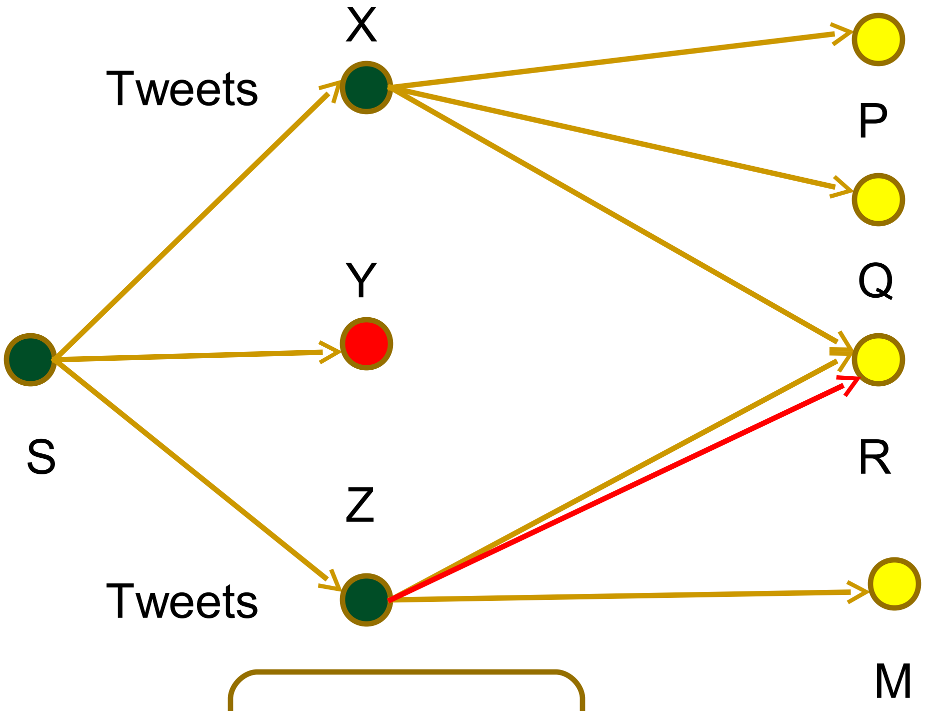
X, Z decide to re-tweet
Y decides not to

Introduction (Continued)

Information Diffusion via Follow & Mention Links

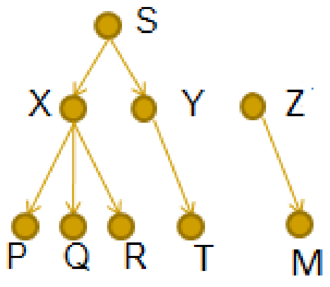
X mentions none

X has 3 followers-
P, Q & R



Z mentions R
Z: … @R

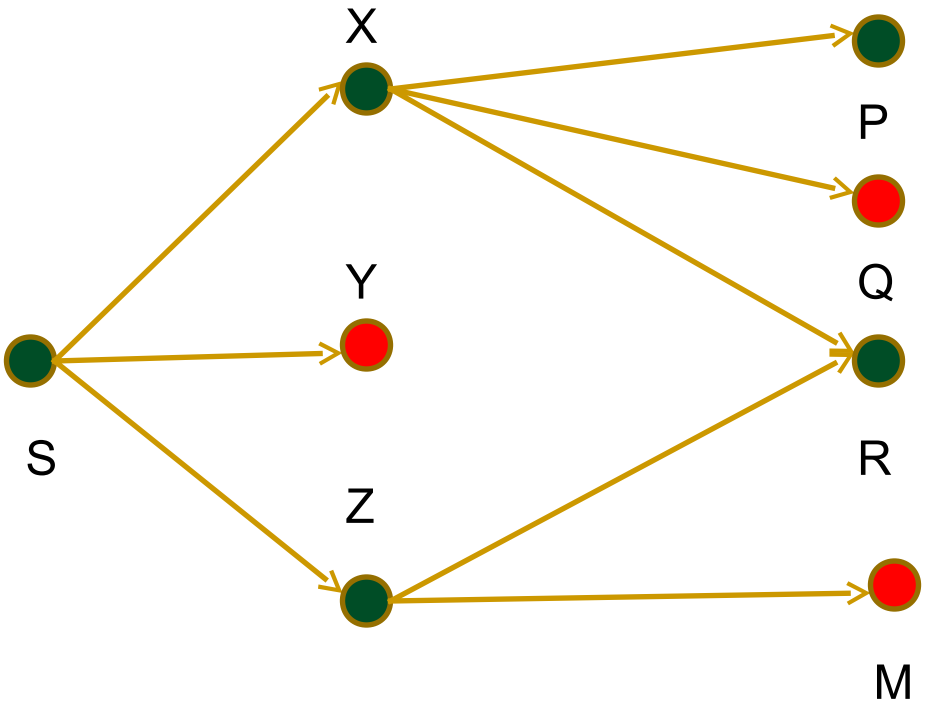
Z has 1 follower M

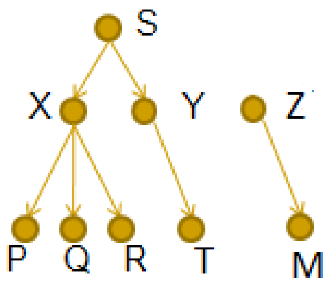


Follower Network

Introduction (Continued)

Information Diffusion via Follow & Mention Links





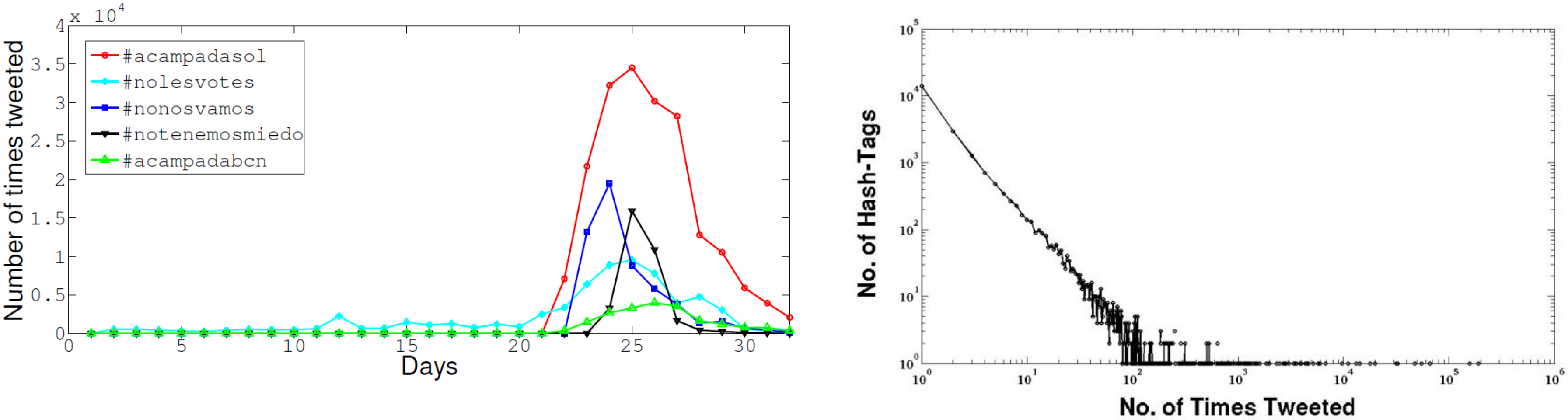
Follower Network

And in this way the information propagates…

P, R decide to propagate
Q, M decide not to propagate

Introduction (Continued)

Hashtag Popularity (Number of users tweeted)



(15m Dataset)

Observation:

1. Different hashtags have different temporal pattern of popularity
2. Few hash-tags are highly popular , but most are not

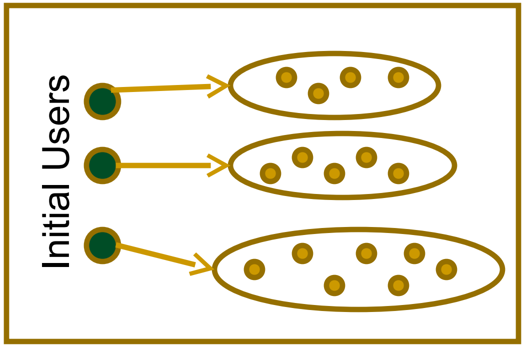
Research Question:

Investigate the Key-factors controlling the popularity of a hashtag

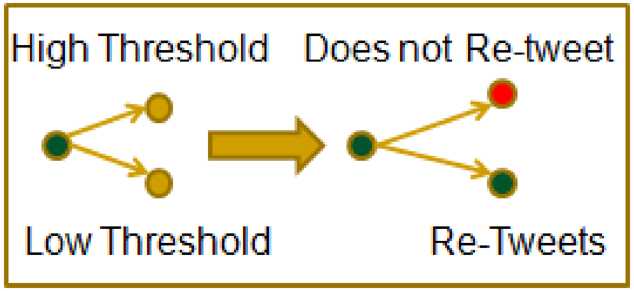
Introduction (Continued)

Factors Influencing Popularity of a hashtag

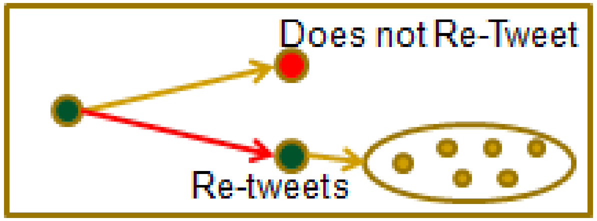
- Number of Initial Users (gets information from external sources)



- Passivity-Threshold of Users

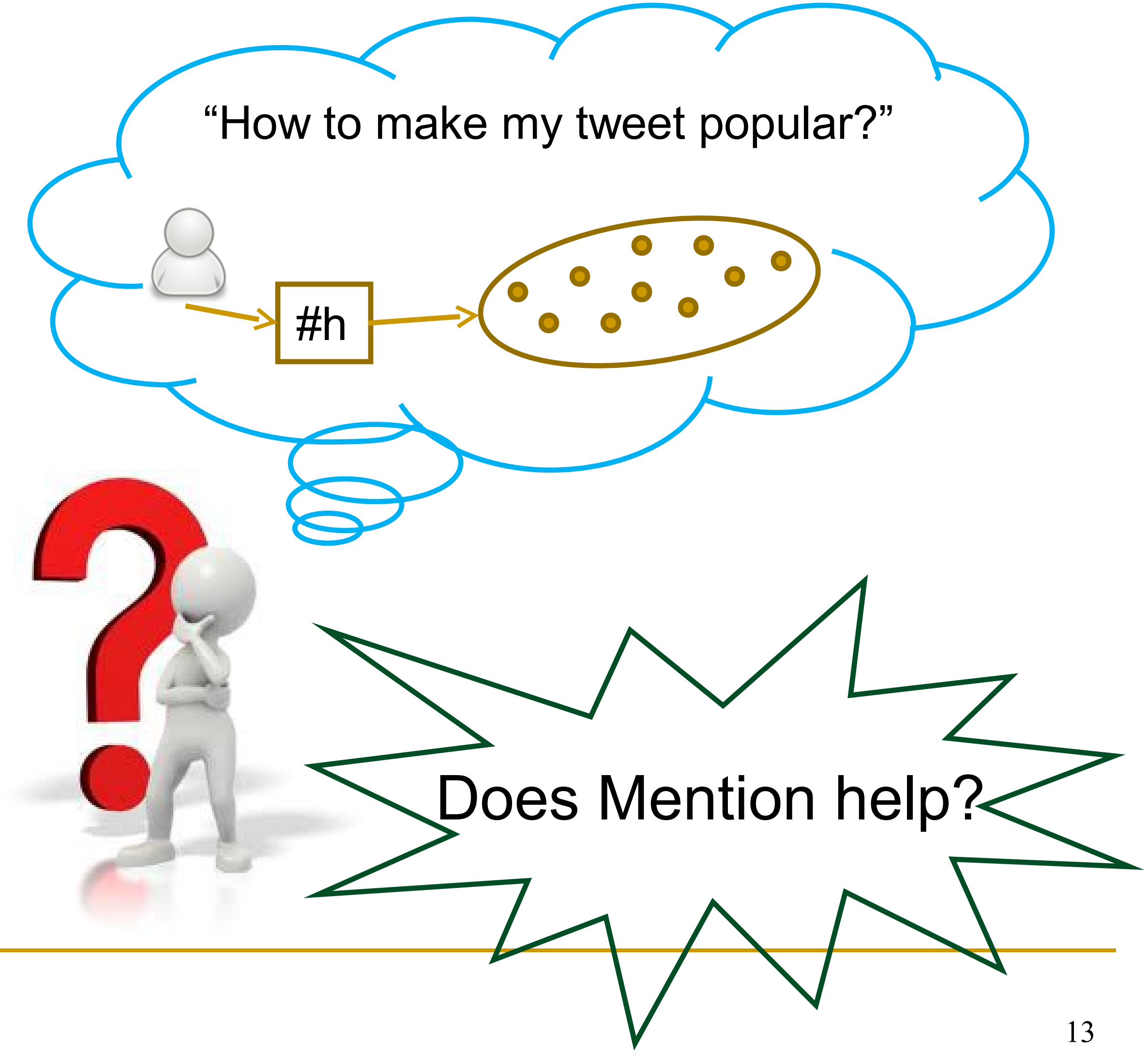


- Mention Usage



Problem Statement

How to increase the popularity of a Hashtag?



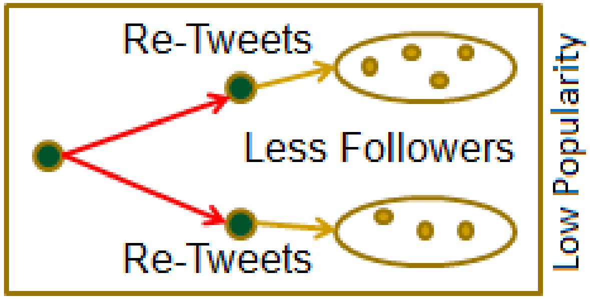
“How to make my tweet popular?”

#h

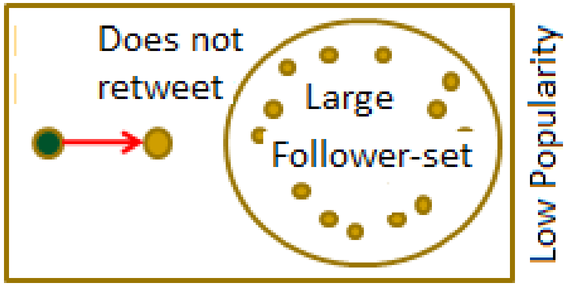
Does Mention help?

Problem Statement (Continued)

Available Options:



- Mention Friends:
  - Pro:
    - High probability of re-tweet
  - Con:
    - Low popularity if friends are not popular (less number of followers).

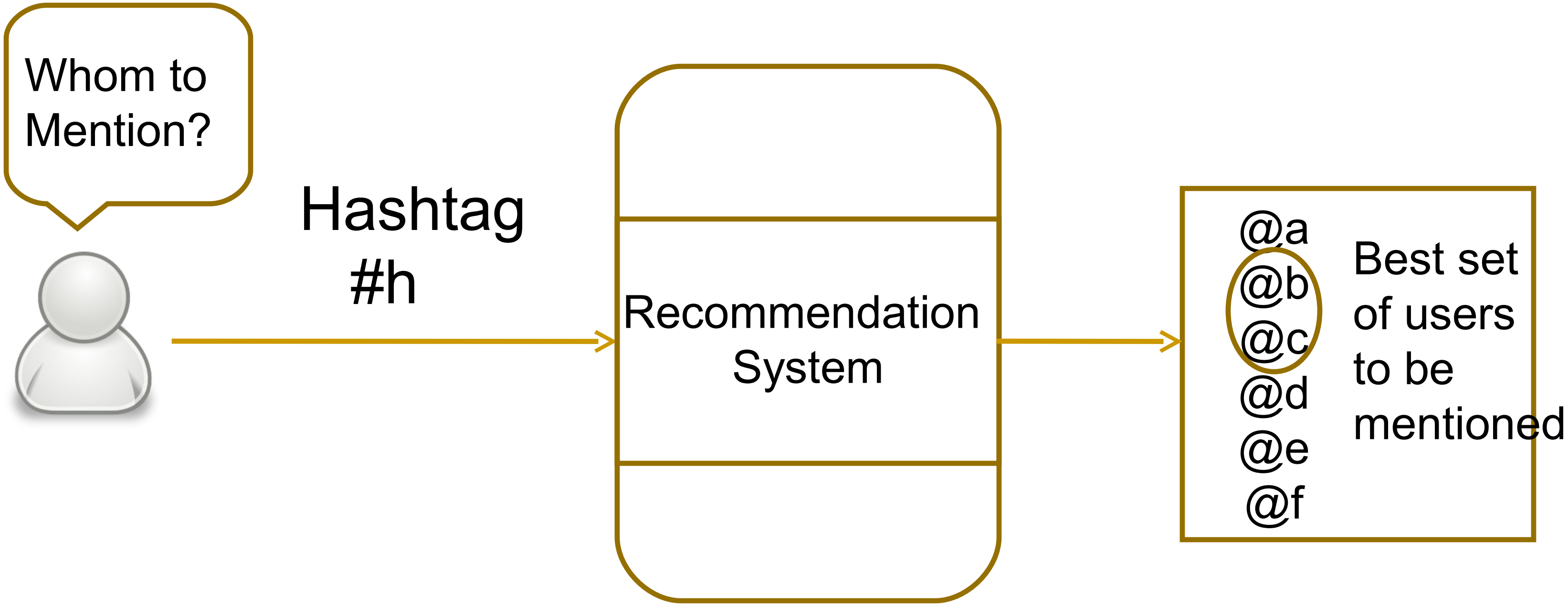


- Mention Celebrities
  - Pro:
    - High popularity if they re-tweet
  - Con:
    - Low probability of re-tweet (high number of followers)

Need to maintain a BALANCE

Objective

Recommendation System



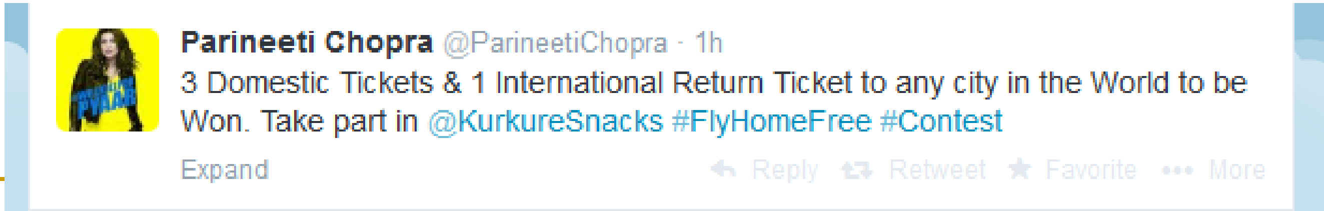
Whom to Mention?

Hashtag #h

Recommendation System

@a
@b
@c
@d
@e
@f

Best set of users to be mentioned



Parineeti Chopra @ParineetiChopra · 1h

3 Domestic Tickets & 1 International Return Ticket to any city in the World to be Won. Take part in @KurkureSnacks #FlyHomeFree #Contest

Outline

- Data-study & Measurements
  - Dataset
  - Representation
  - Dependency on Mention
- Model for hashtag Propagation
  - Description
  - Set Parameters from Dataset
  - Validation
  - Insights
- Conclusion

Data-study & Measurements

Dataset

- 15m Dataset ➔ Contains information about tweets posted during the revolutionary movements in Spain during May 2011

- Data:
  - user_id ; timestamp ; hashtag_list ; mention_list
  - Follow Links

- Statistics:
  - Total users: 87569
  - Total Tweets: 529393
  - Hash-Tags: 22376

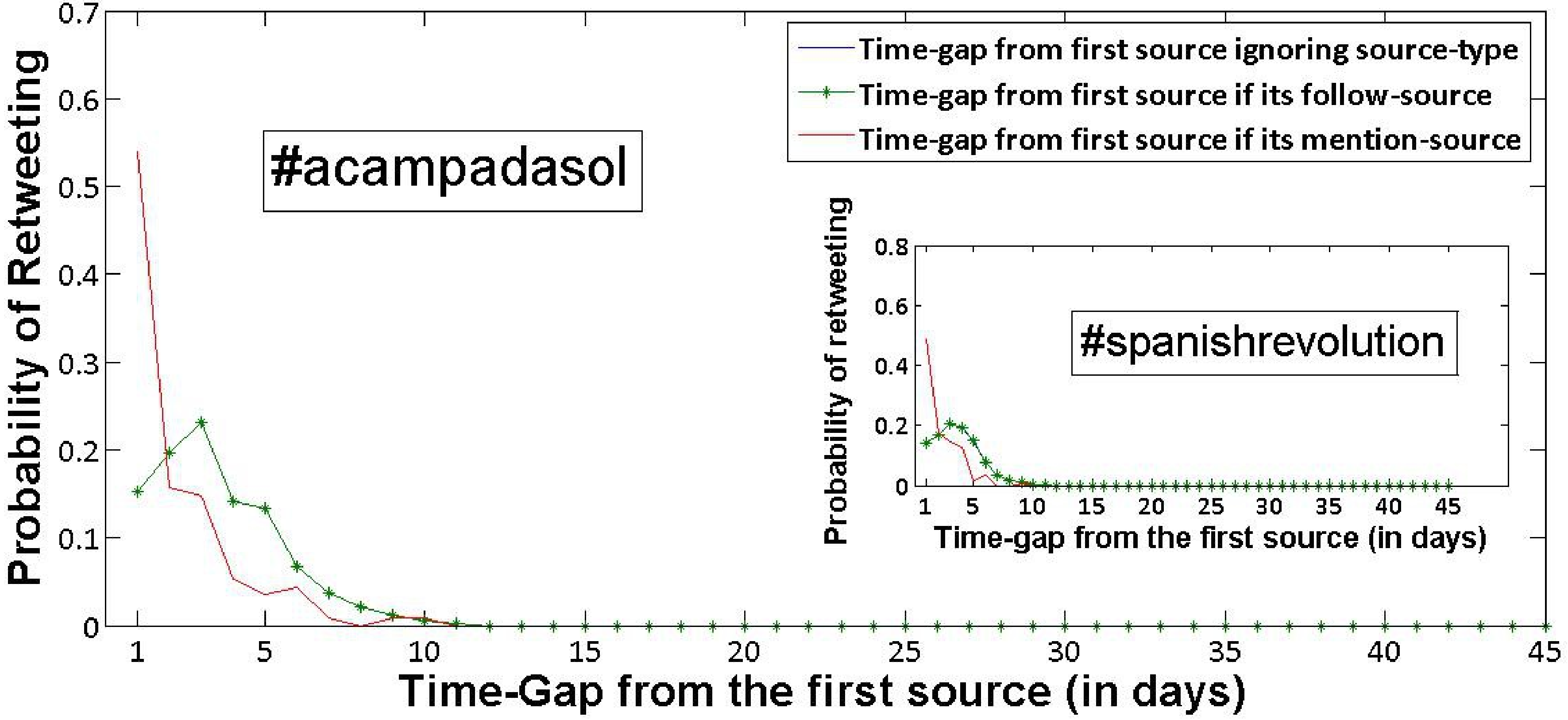
Filtered relationships only to those who sent at least a message in that **topic**; or were mentioned by someone who did.

Source: *http://cosnet.bifi.es/research-lines/online-social-systems/15m-dataset*

Y. Moreno et al. The dynamics of protest recruitment through an online network. Scientic reports, 1, 2011

Other datasets: Arab-spring datasets

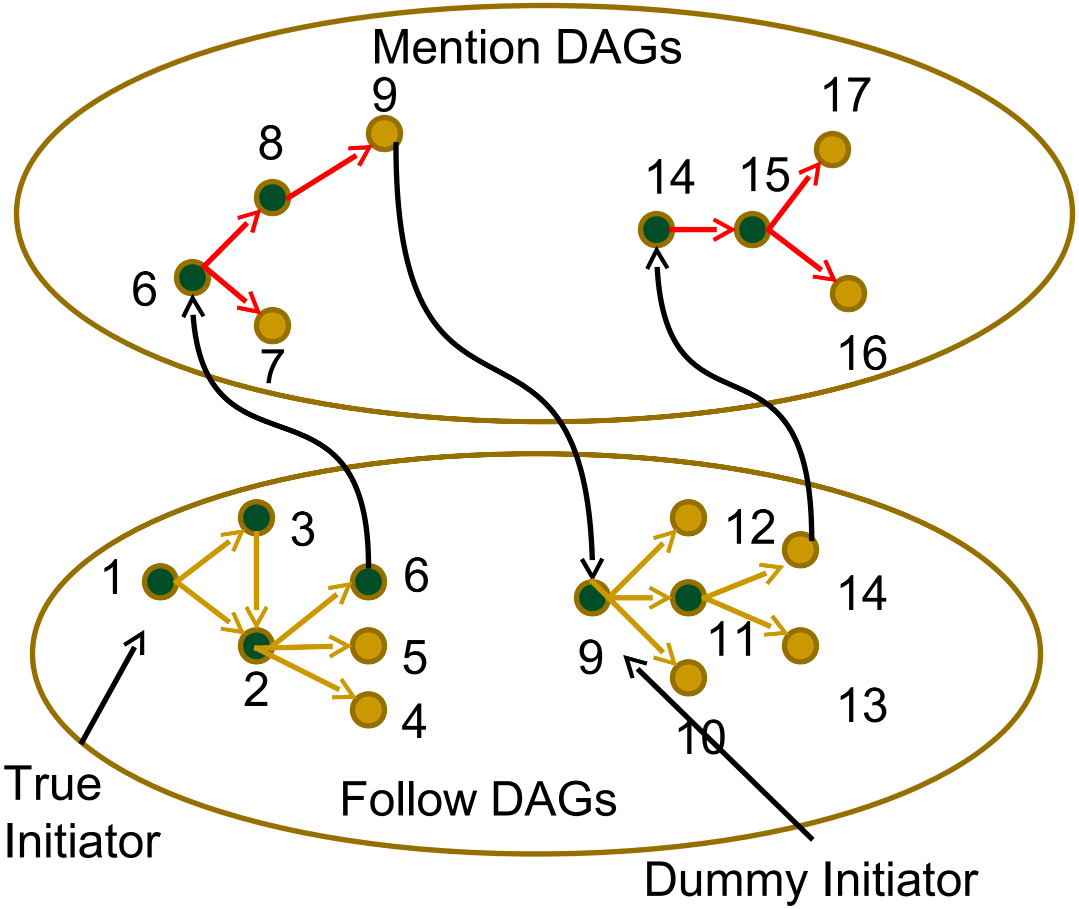
Data-study & Measurements



Retweet comes faster if the source is “Mention”

Data-study & Measurements

Representation



True Initiators

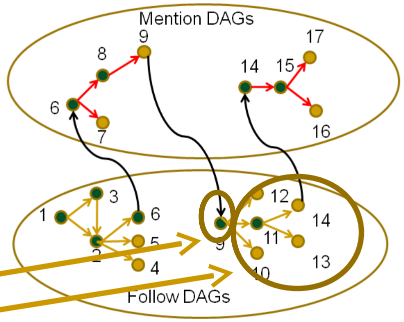
Those who get the information from external Source and spread in the network
e.g. 1

Dummy Initiators

Those who get the Information only through Mention links and spread In the network
e.g. 9

Data-study & Measurement

Dependency on Mention



- For each hashtag ‘#h’,
  - Set of dummy initiators (set A)
  - Set of users who belong to only the DAGs rooted by dummy initiators (set B)
  - Set of users who have tweeted #h (set C)

- The users in sets A & B would not have got the information (‘#h’) without “Mention”

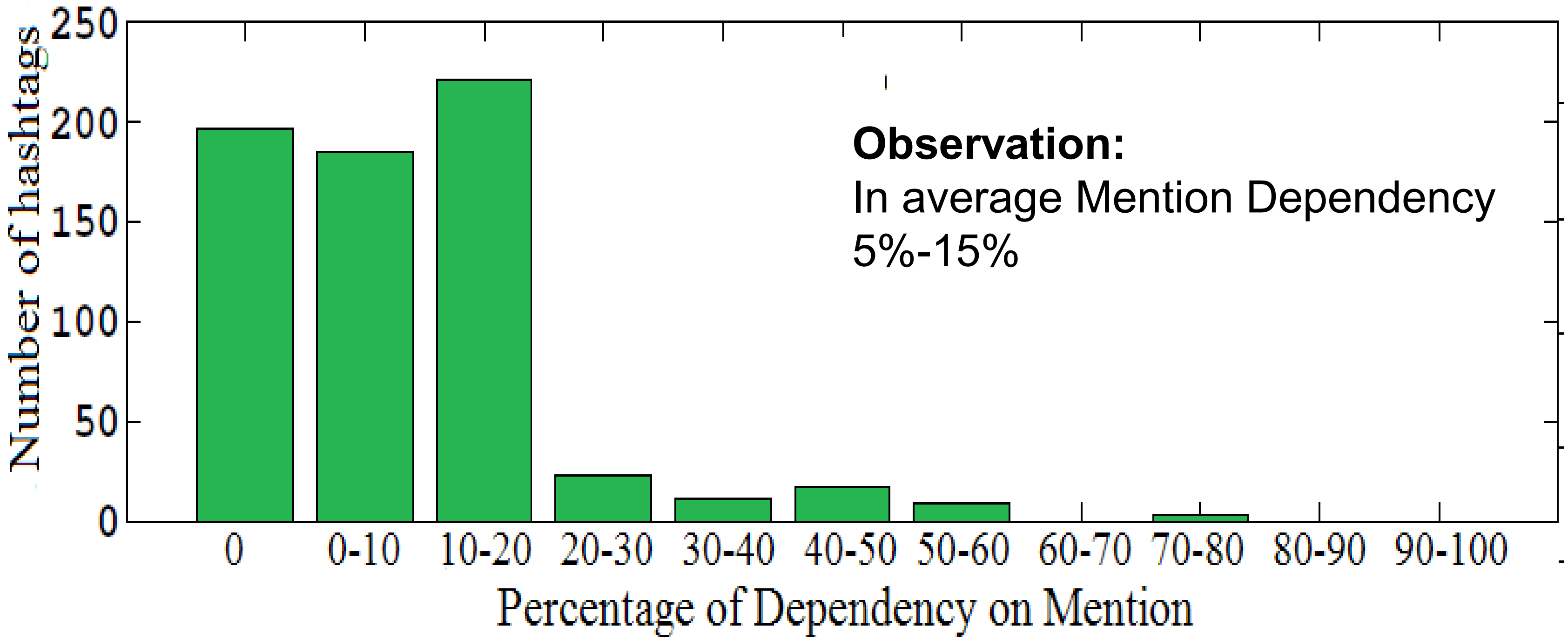
- Dependency of ‘#h’ on Mention =

  Fraction of tweeting users who would not have received the hashtag without Mention

  = (|A|+|B|)/|C|

Data-study & Measurements

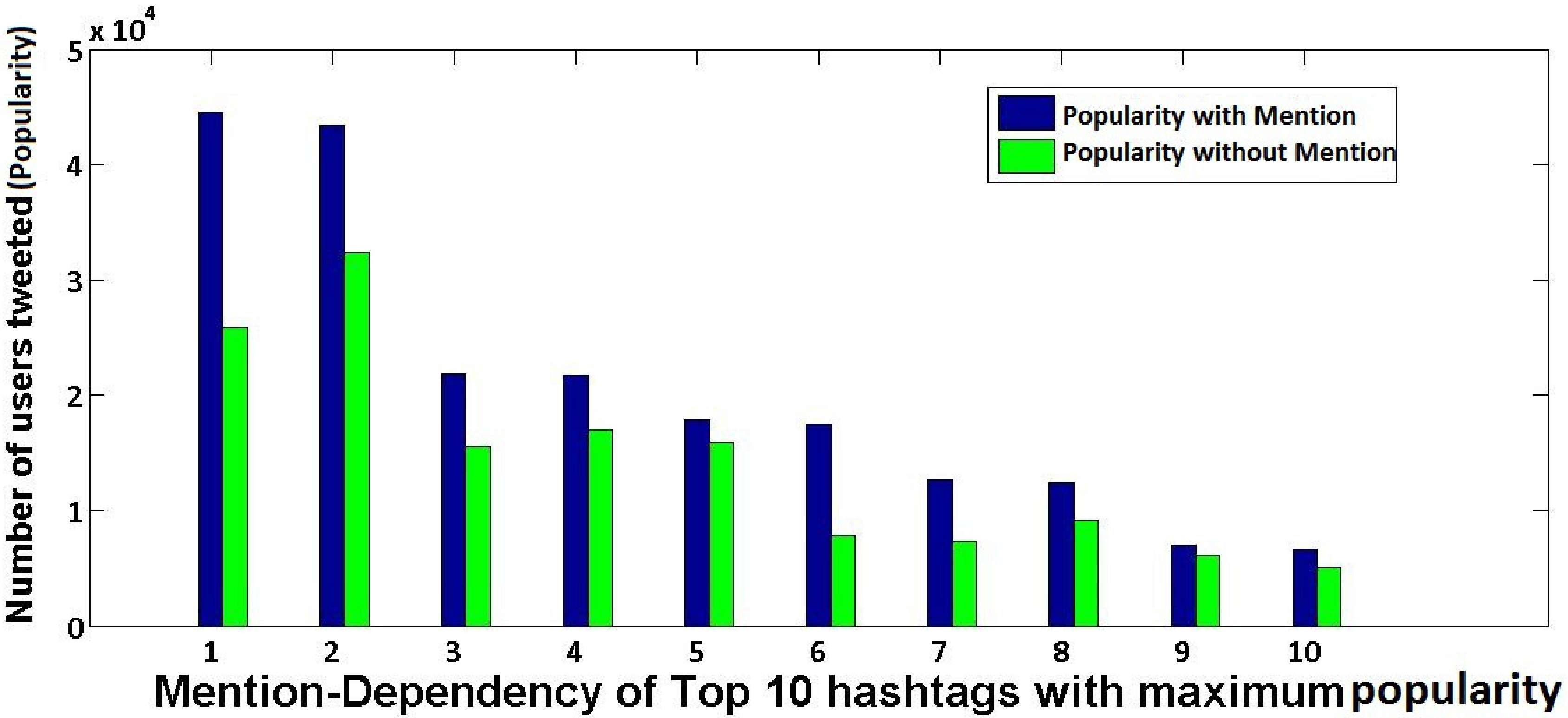
Dependency on Mention



Observation:
In average Mention Dependency
5%-15%

Data-study & Measurements

Dependency on Mention



Data-study & Measurements

Group Hashtags based on Dependency on Mention

- Grouped the hash-tags based on their dependency on Mention
  - Group 1: Consists of top 41 Mention-dependent hashtags

    Average popularity: 8092.7

    Average number of mentioned users: 484.83

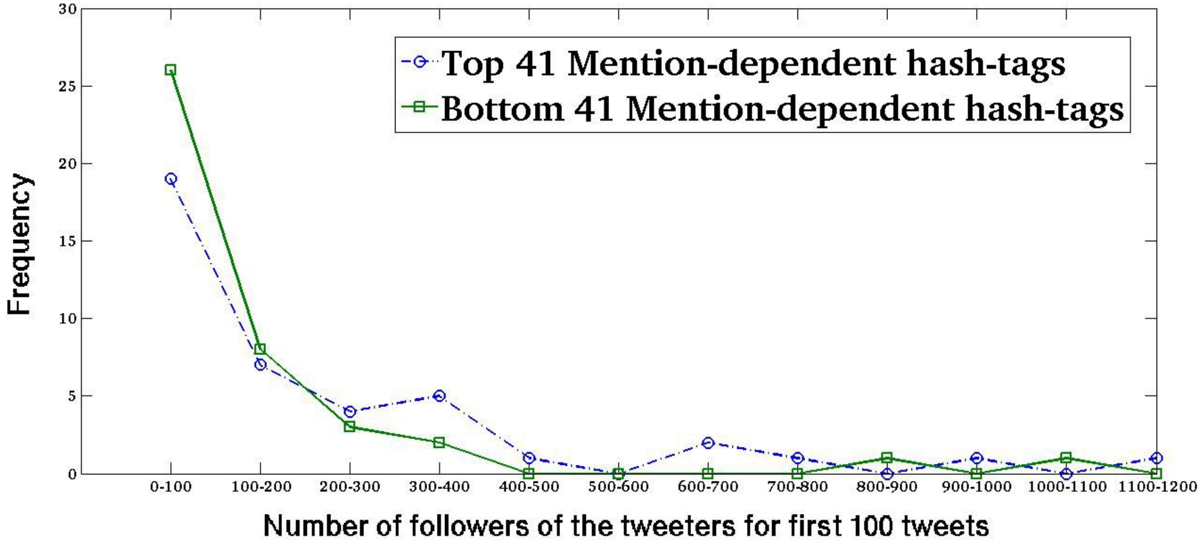
  - Group 2: Consists of bottom 41 Mention-dependent hashtags

    Average popularity: 82.21

    Average number of mentioned users: 4.48

- Observation: Highly popular hashtags are generally more dependent on "Mention"

What did the tweeters of Group 1 hashtags do differently from the tweeters of Group 2 hashtags in the first 100 tweets?

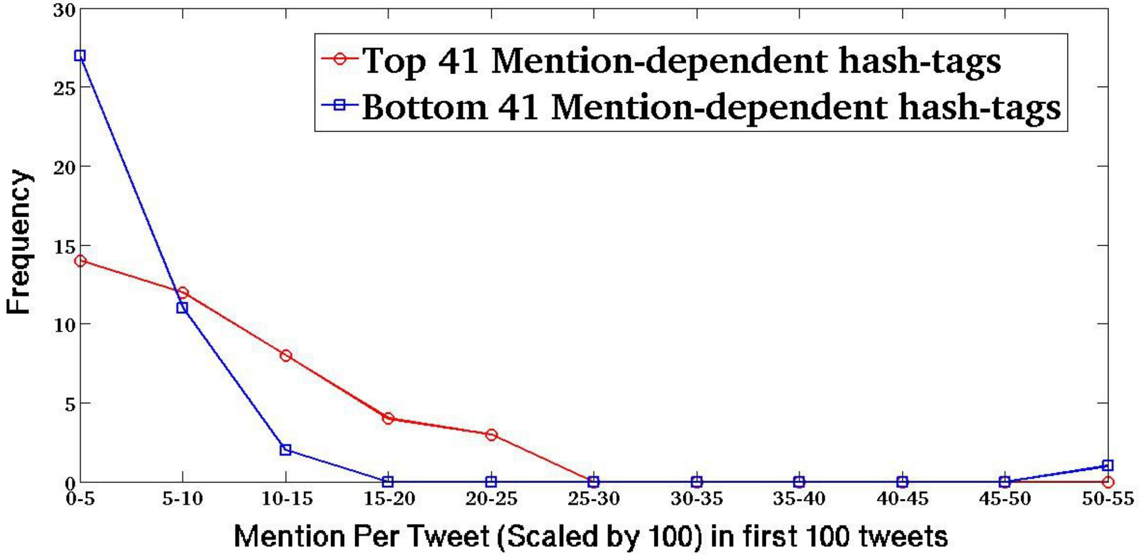


Observation:

Almost similar popularity of initial users

Observation:

Group 1 hashtag tweeters used mention at almost double of the rate of Group 2 tweeters



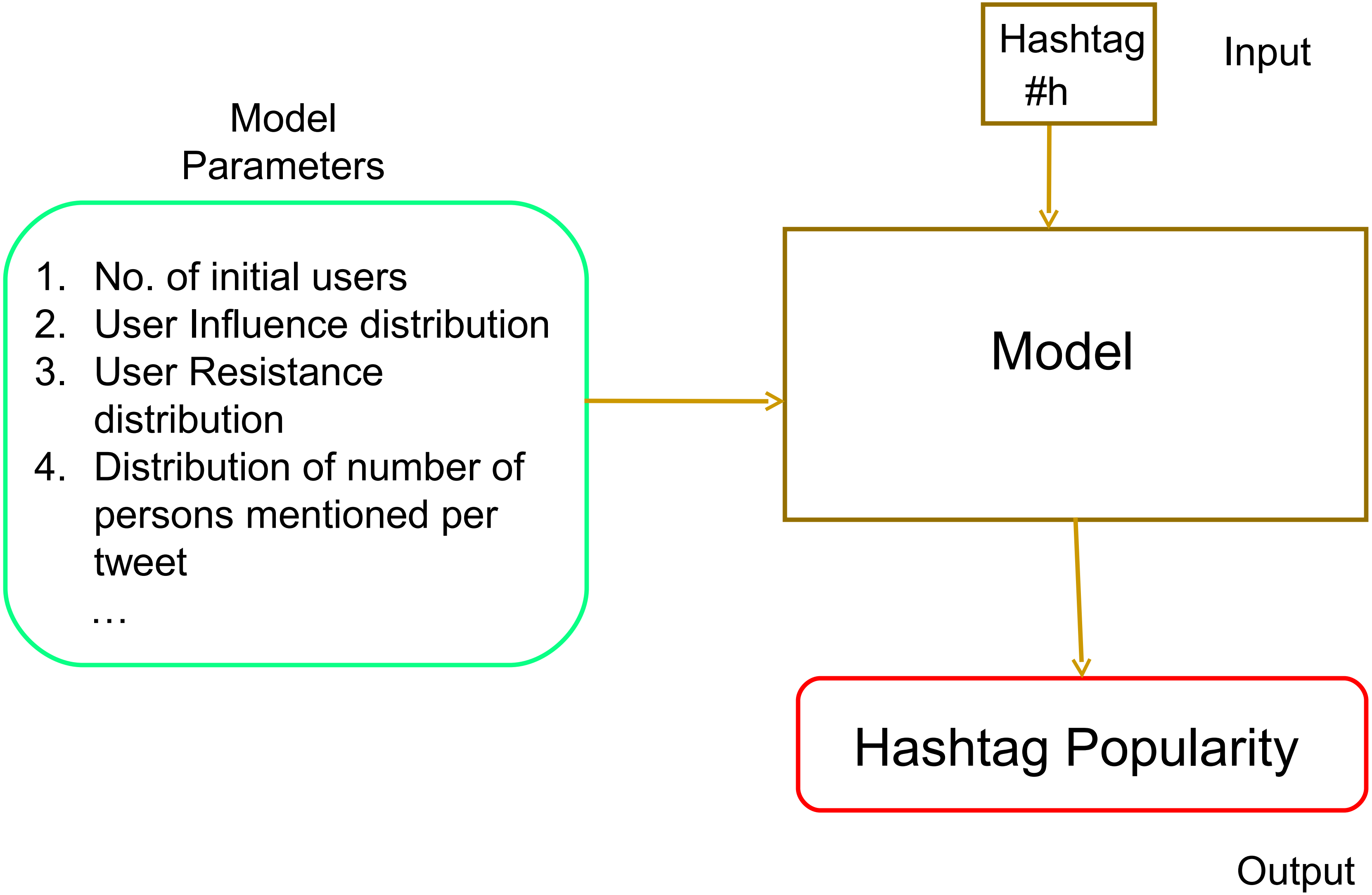
Outline

- Data-study & Measurements
  - Dataset
  - Representation
  - Dependency on Mention

- Model for hashtag Propagation
  - Description
  - Set Parameters from Dataset
  - Validation

- Conclusion

Model for Hashtag Propagation



Hashtag #h

Input

Model Parameters

1. No. of initial users
2. User Influence distribution
3. User Resistance distribution
4. Distribution of number of persons mentioned per tweet

…

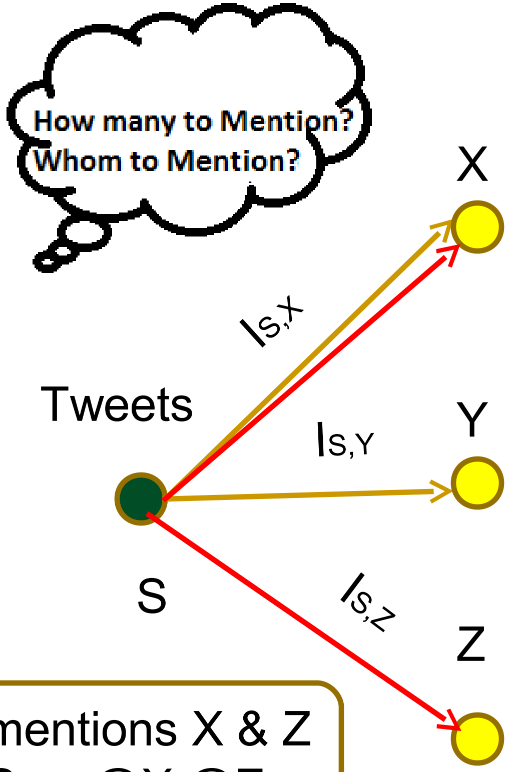
Model

Hashtag Popularity

Output

Model for Hashtag Propagation

Intuition Behind the Model

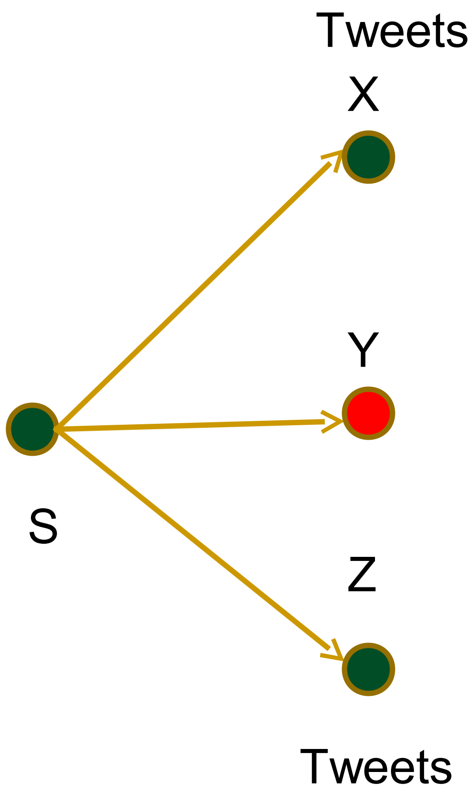


S mentions X & Z
S: …@X @Z

S has two followers X & Y

Model for Hashtag Propagation

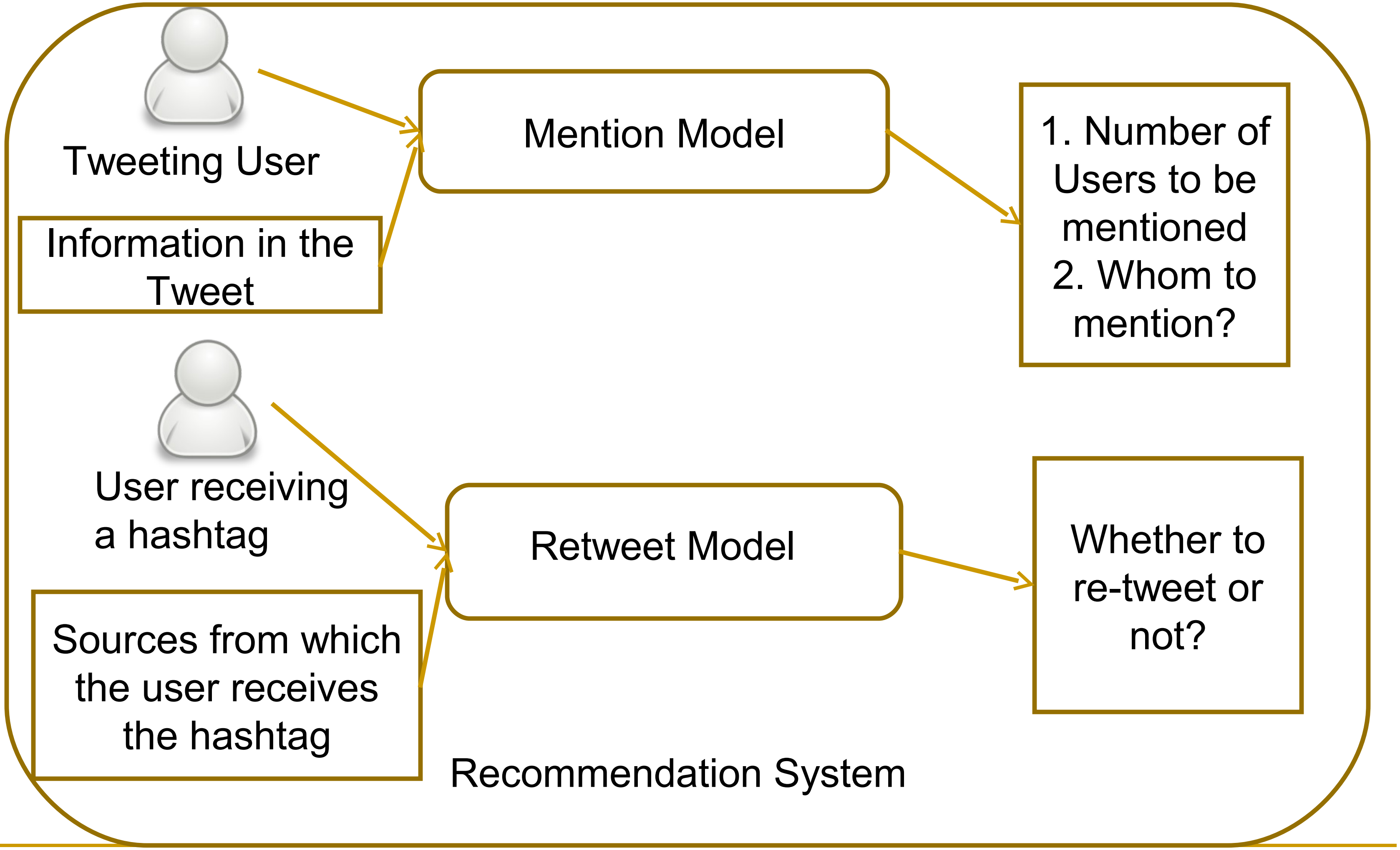
Intuition Behind the Model



X, Z decide to propagate
Y decides not to propagate

Model for Hashtag Propagation

Model Components



Tweeting User

Information in the Tweet

Mention Model

1. Number of Users to be mentioned
2. Whom to mention?

User receiving a hashtag

Sources from which the user receives the hashtag

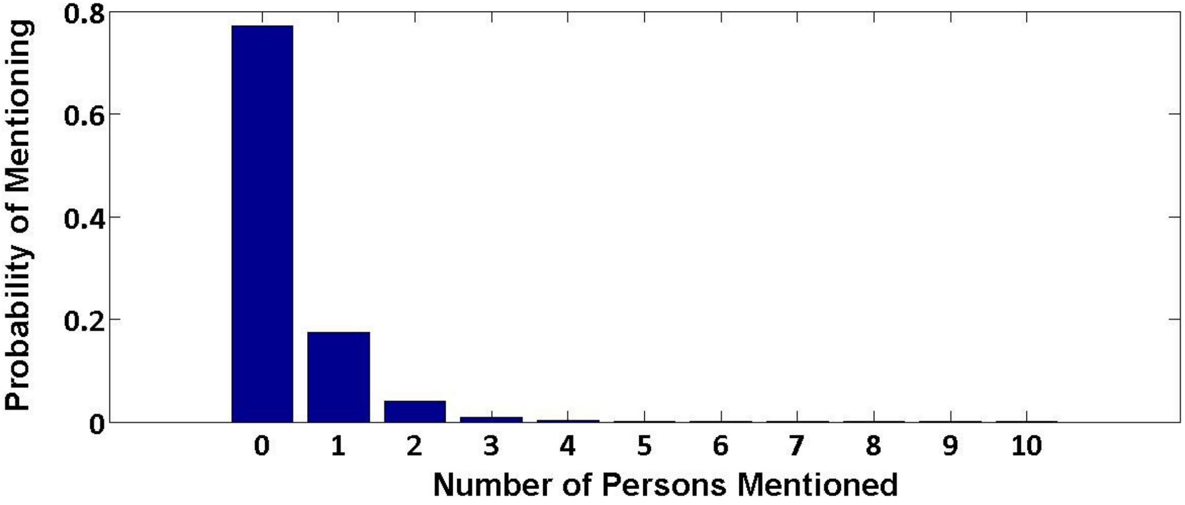
Retweet Model

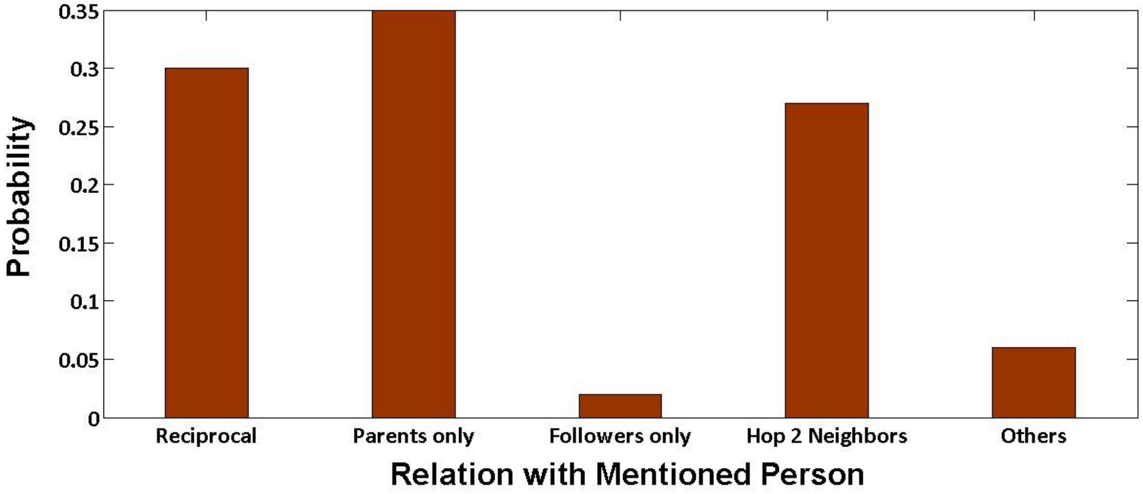
Whether to re-tweet or not?

Recommendation System

Mention Model

- Distributions calculated from the Dataset





Retweet Model (Linear Threshold Based)

- We calculate the weightage of each link a user (say,u1) gets
  - Factors:

- Influence of the user (say, u2) from whom the link is coming (Calculated using PageRank)
- Importance of the link based on the
  - Type of Link (Follow/Mention/Mixed)
  - Social Tie between u1 & u2 (Reciprocity) (calculated from dataset)
- Time-gap between u2’s tweet and the current time

- We also calculate the passivity/resistance of each user

$$1 - \frac{\text{Number of times tweeted}}{\text{Number of times received any hashtag}}$$

Retweet Model (M/C Learning Based)

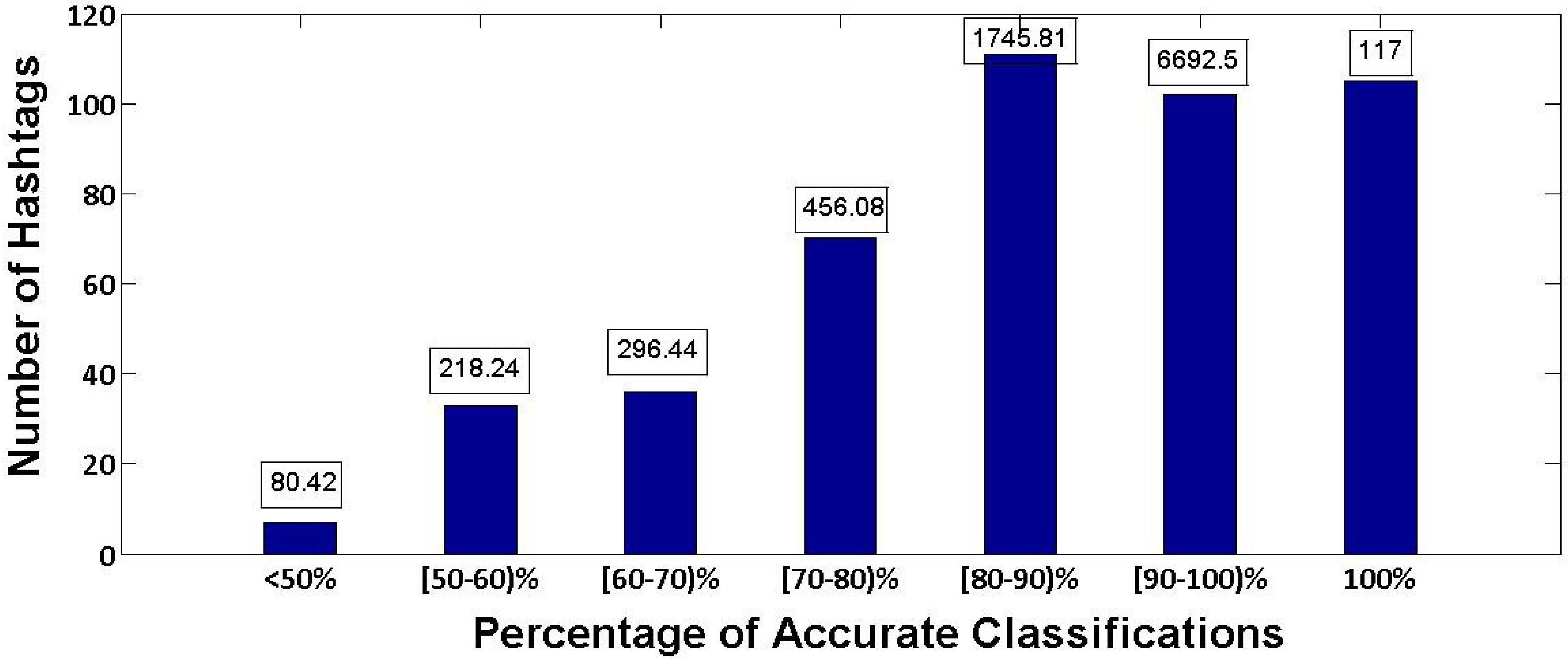
- For each Retweet,
  - Collect the following features of each of the sources
    - Influence using PageRank
    - Number of followers
    - Time-Gap
    - Type of Link (Follow/Mention)
    - Relation with retweeting user (follower/parent/reciprocal)
  - Combine the features by taking "Average" and "Standard Deviation"
  - Collect the following features of the retweeting user
    - Influence using PageRank
    - Number of followers
    - Number of times tweeted
    - Number of times tweeted the current hashtag
    - Number of times got exposed to the current hashtag
  - Total 16 features per retweet (including "Number of sources")

Retweet Model (M/C Learning Based)

- Collect the features for
  - Positive Class: Users who received the information and retweeted
  - Negative Class: Users who received the information but did not retweet
- Train a Support Vector Machine using those feature-vectors

Retweet Model (M/C Learning Based)

$$\text{Accuracy} = \frac{|\text{Correctly classified samples}|}{|\text{Test Set}|}$$



Validation (Using threshold based Retweet Model)

1. Simulate model with a fixed number of initial users with different parameter values and get the popularity values

2. From dataset, collect hashtags with almost same number of initial users as the simulation

3. Check whether popularities of those real hashtags from the dataset fall within the range of simulated popularity values

Hashtag	Real Original Tweeters	Real Final Popularity	Predicted Popularity for 1100 initial users
#worldrevolution	1091	1866	[1341-3733]
#acampadasol	1125	3449	[1341-3733]

Outline

- Data-study & Measurements
  - Dataset
  - Representation
  - Dependency on Mention
- Model for hashtag Propagation
  - Description
  - Set Parameters from Dataset
  - Validation
- Conclusion

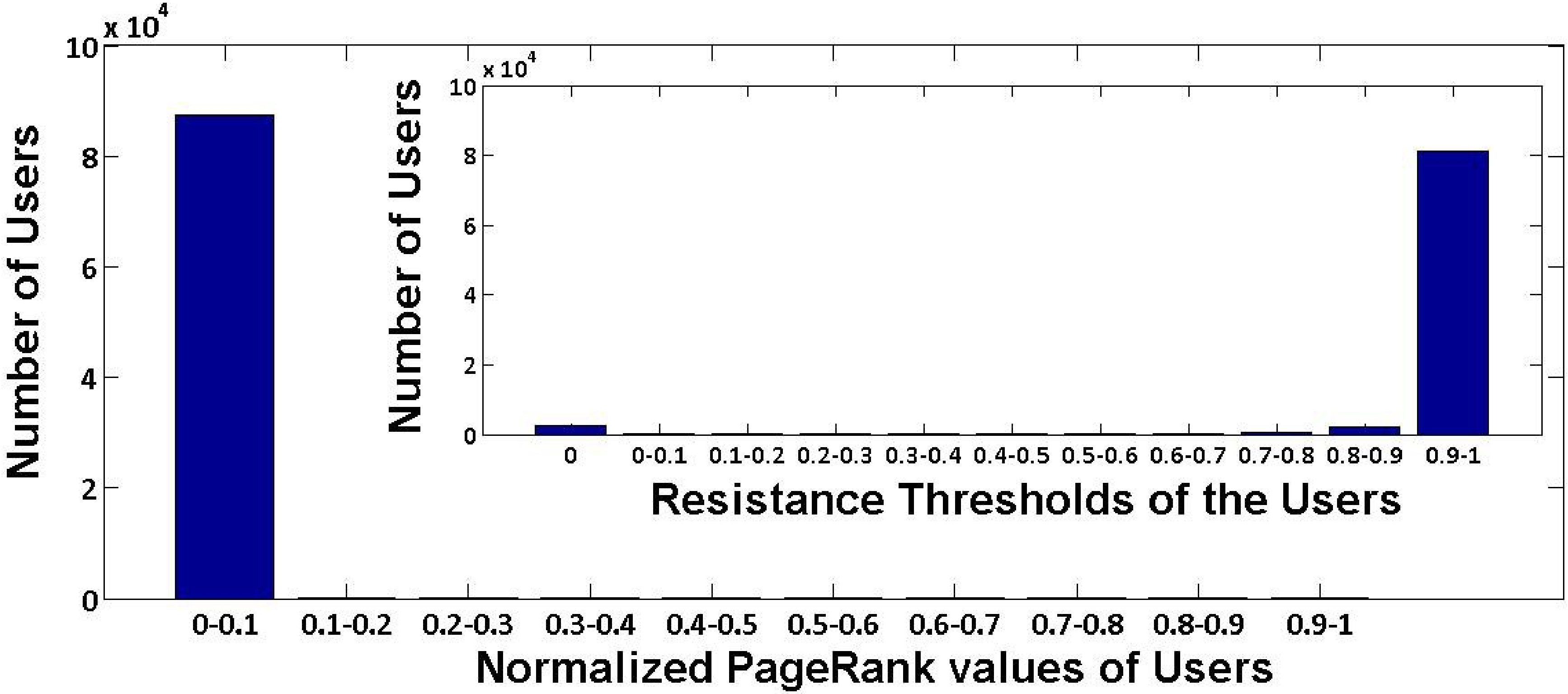
Conclusion

- “Mention” definitely plays a key-role in deciding the
  - Popularity of hashtags
  - Speed at which the hashtag propagates
- Using insights from simulations of our model, our recommendation System should try to
  - Suggest minimum number of users (due to character limitation of tweets)
  - So that maximum popularity can be achieved within less time

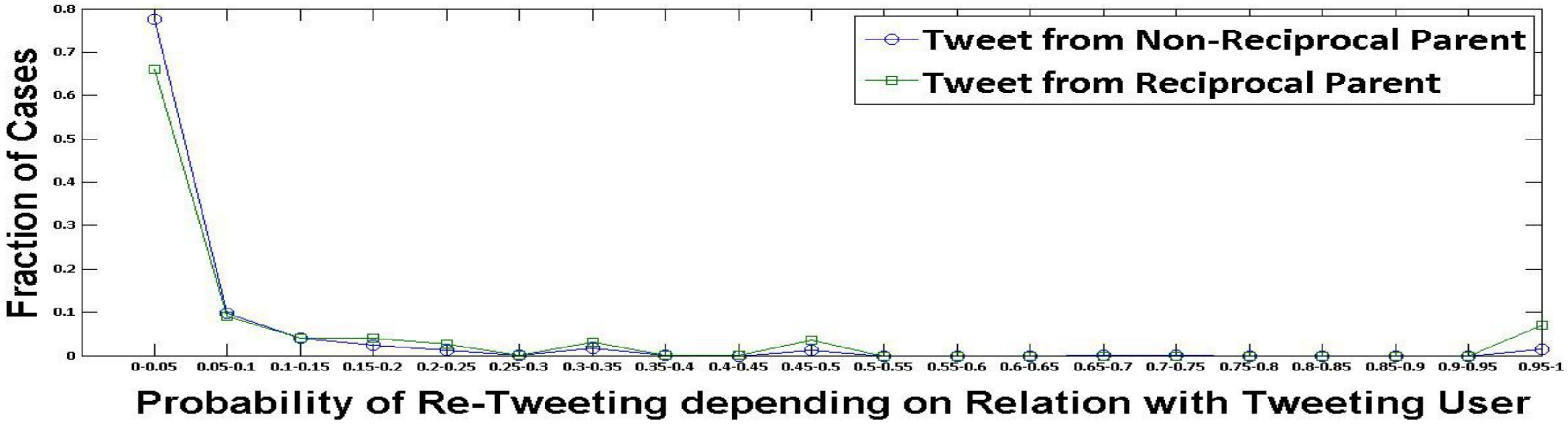
Thank You

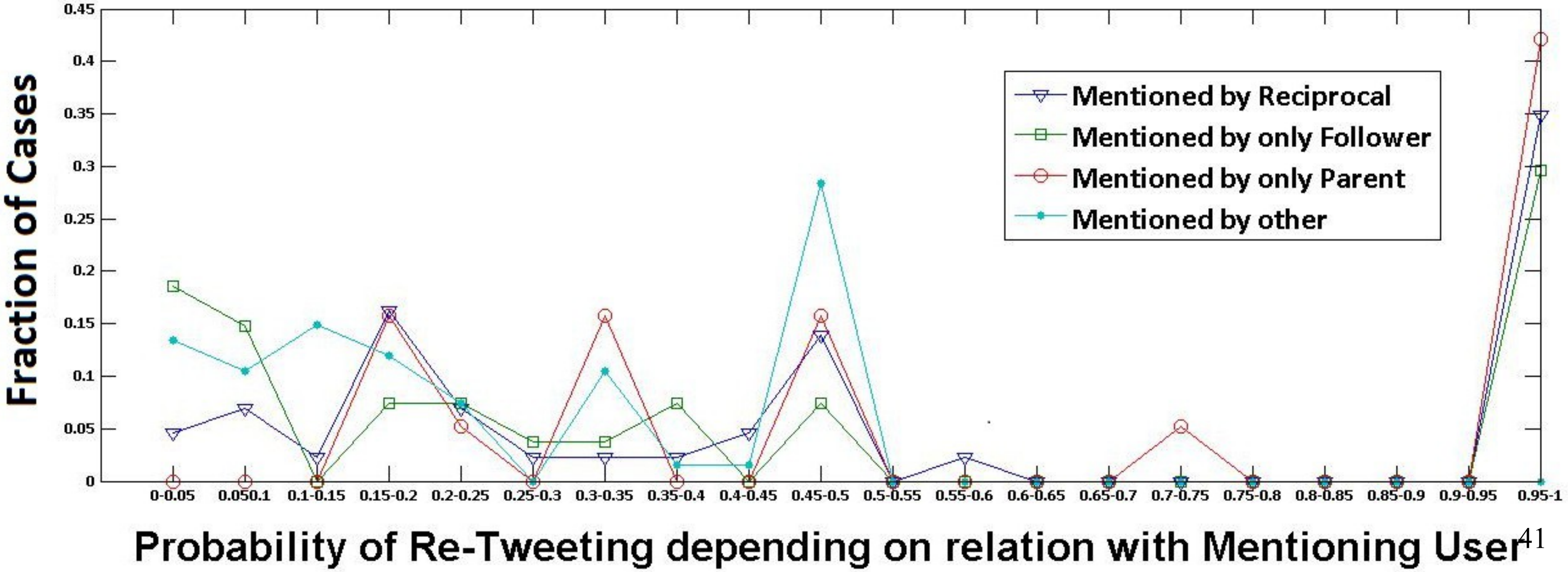
- BACK-UP

User Influence & User Resistance Distribution



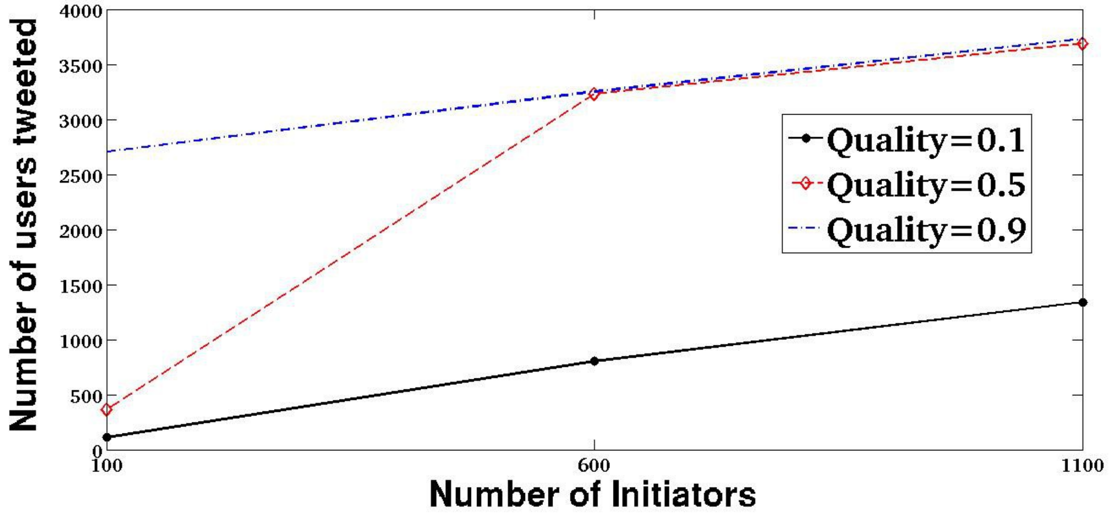
Importance of Type of Links



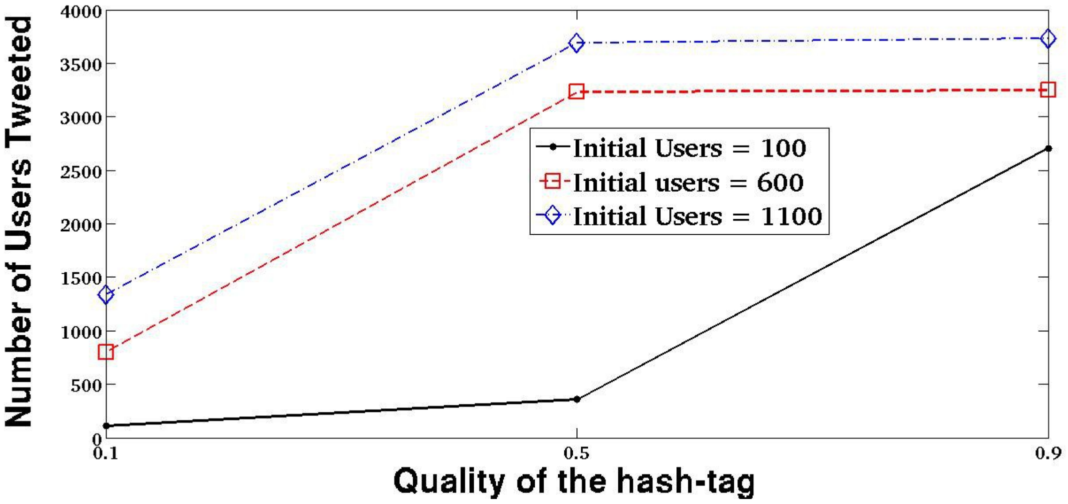


- Simulation : Vary the number of initial users
  Vary the quality of the hashtag

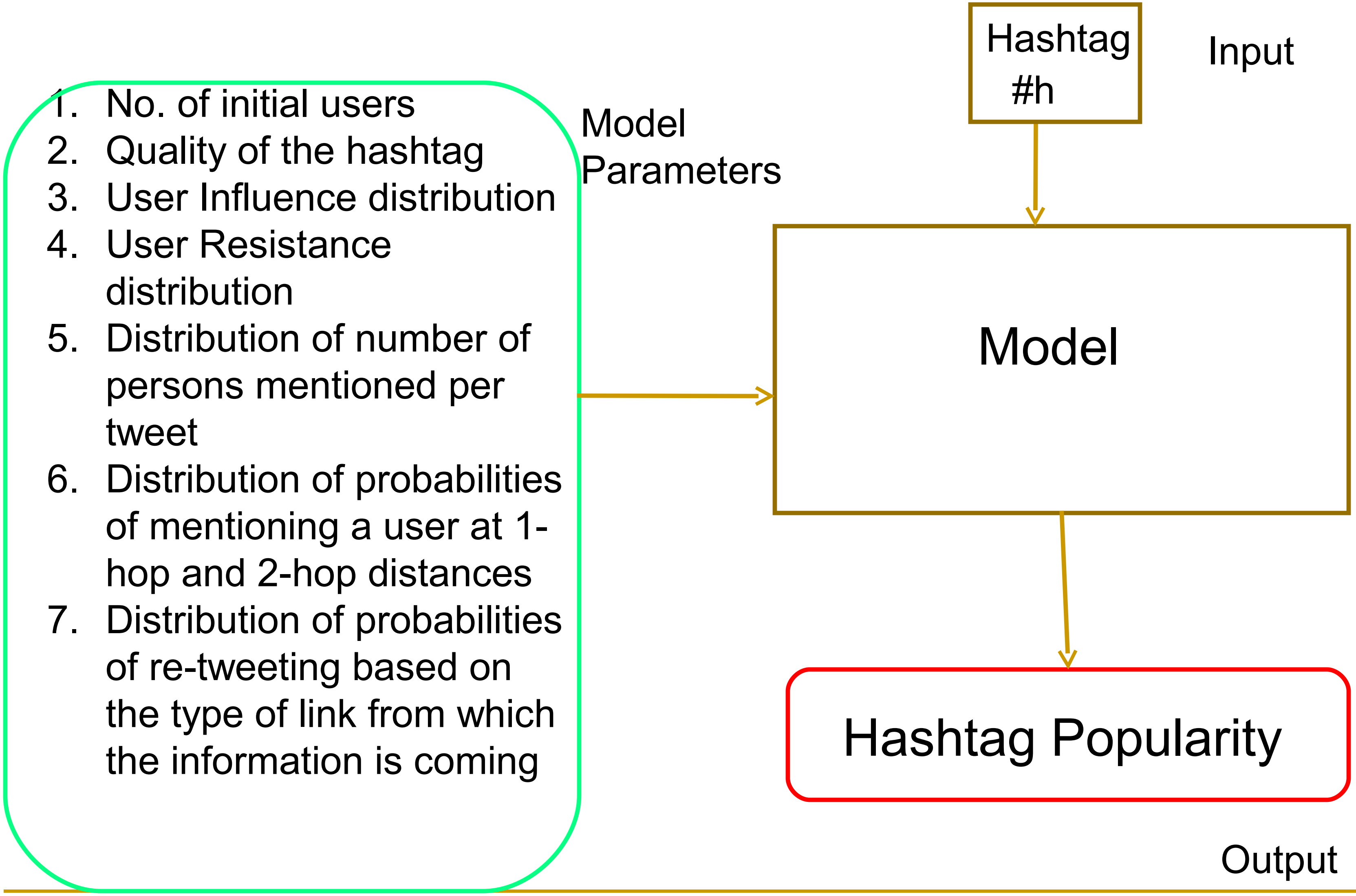
(set α=0.5, β=0.5. Calculate other parameters from dataset)



Next, we will validate whether we are getting similar reach for **real hashtags** with similar number of original tweeters.



Model for Hashtag Propagation



1. No. of initial users
2. Quality of the hashtag
3. User Influence distribution
4. User Resistance distribution
5. Distribution of number of persons mentioned per tweet
6. Distribution of probabilities of mentioning a user at 1-hop and 2-hop distances
7. Distribution of probabilities of re-tweeting based on the type of link from which the information is coming

Model Parameters

Hashtag #h

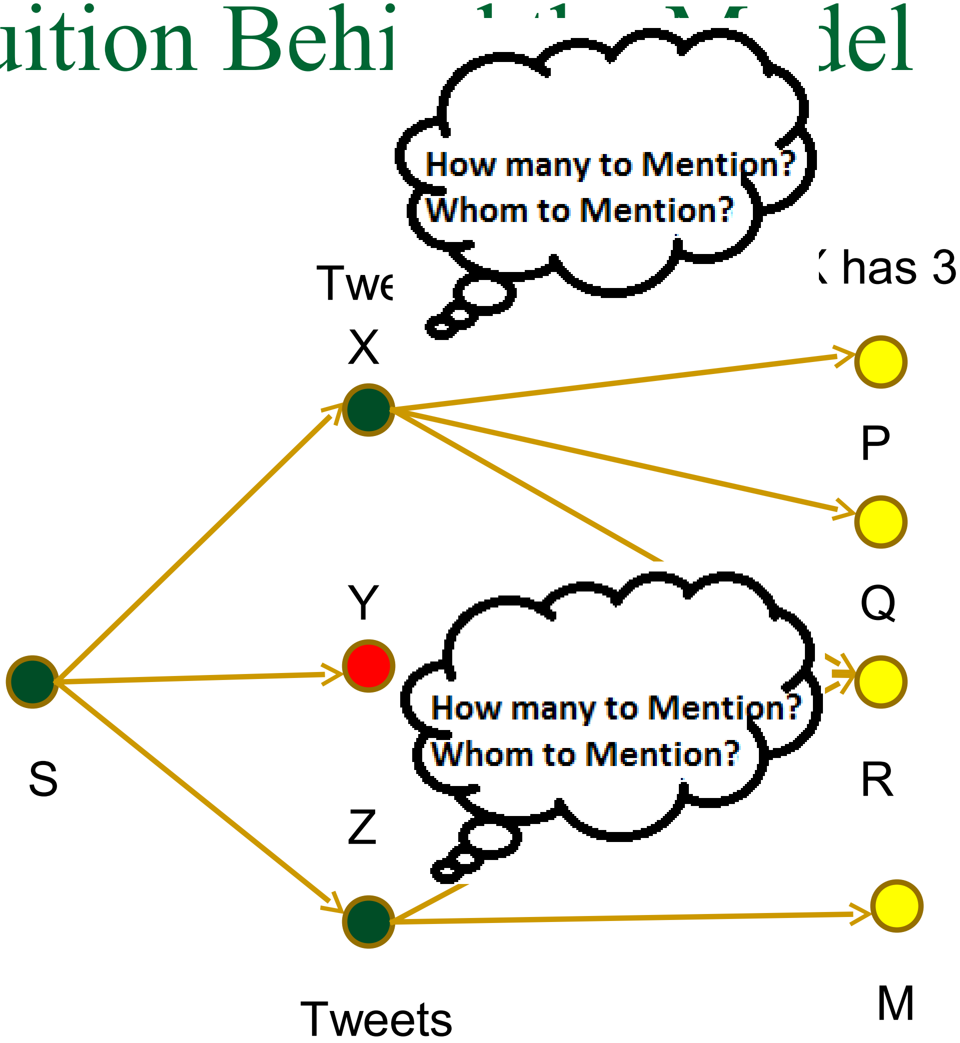
Input

Model

Hashtag Popularity

Output

Intuition Behind the Model

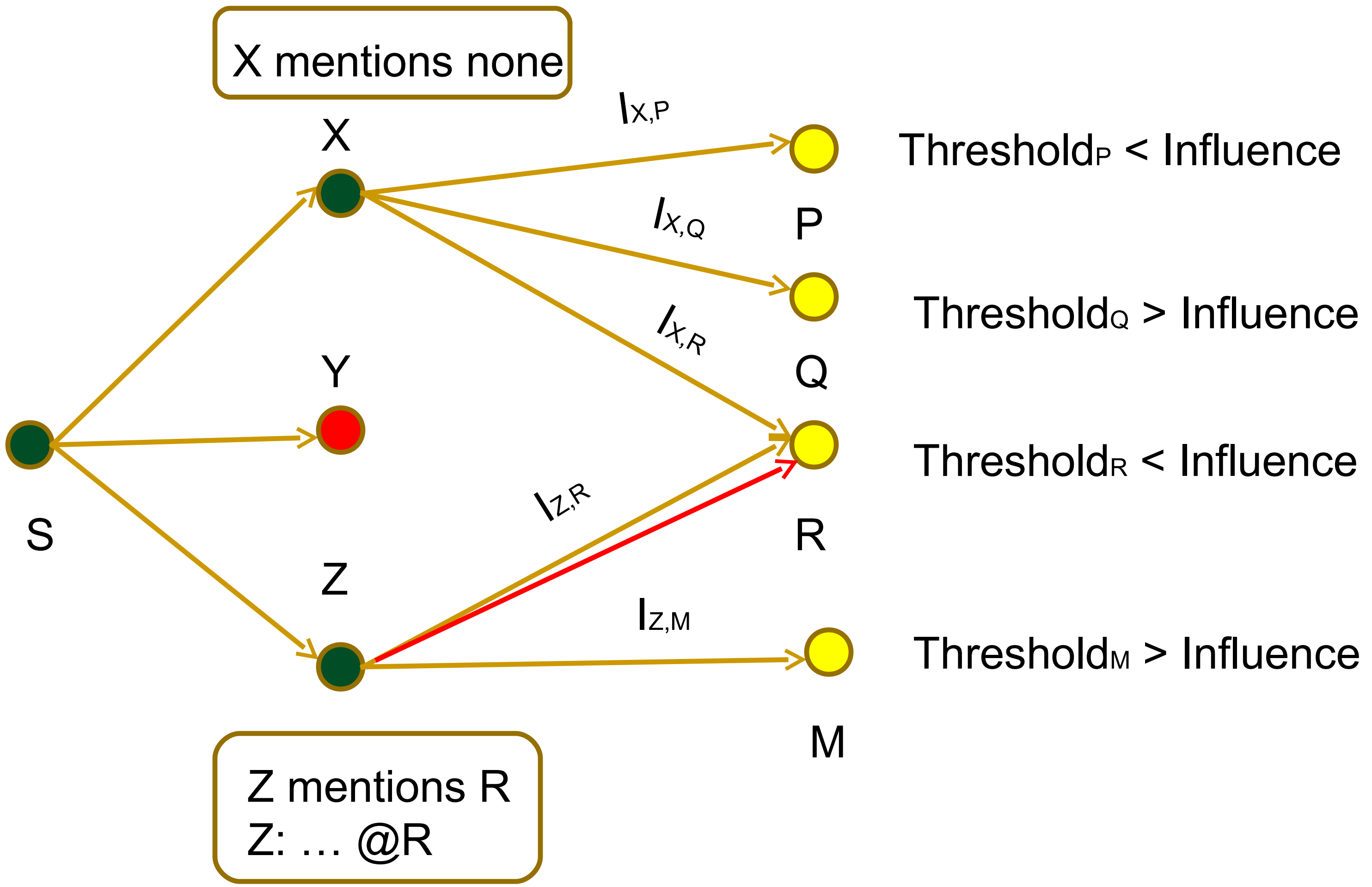


X has 3 followers- P, Q & R

X, Z decide to propagate
Y decides not to propagate

Z has 1 follower M

Intuition Behind the Model



X mentions none

Z mentions R
Z: … @R

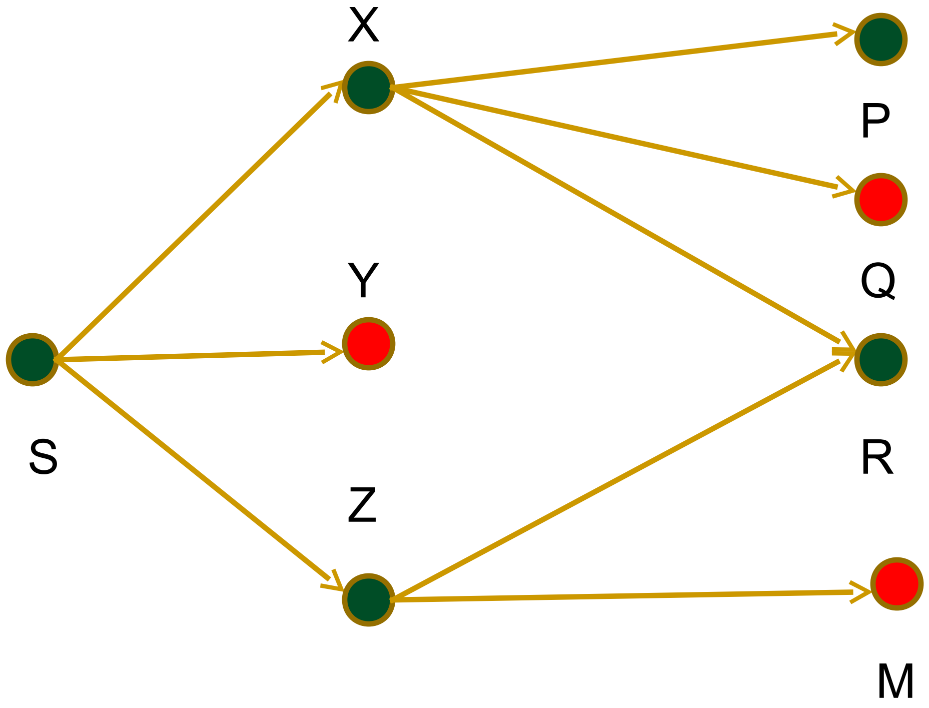
Threshold$_P$ < Influence

Threshold$_Q$ > Influence

Threshold$_R$ < Influence

Threshold$_M$ > Influence

Intuition Behind the Model



P, R decide to propagate
Q, M decide not to propagate

And in this way the information propagates…

Validation

Simulation results for, Initial Users= 1100

1340.77	3032.6	3689.89	3729.53	3733.26

Popularity for different set of parameters

Hash-tag	Original Tweeters	Final Reach
#worldrevolution	1091	1866
#acampadasol	1125	3449

Simulation results for, Initial Users= 100

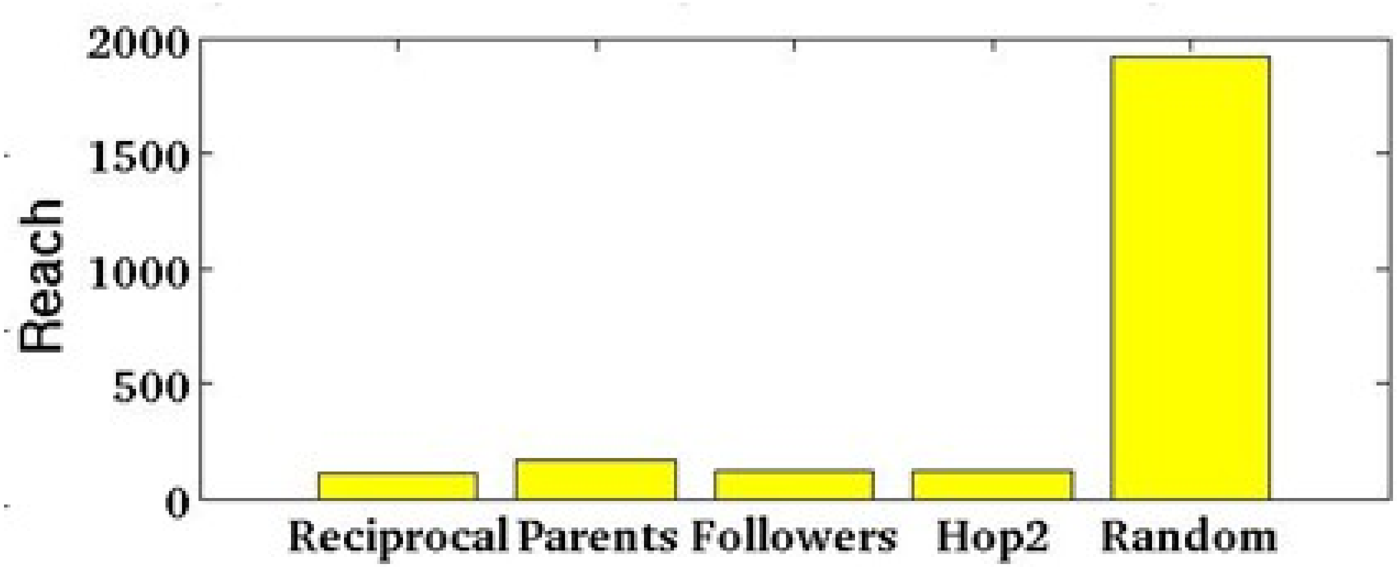
111.23	158.266	359.43	2179.03	2705.466

Popularity for different set of parameters

Hash-tag	Original Tweeters	Final Reach
#upyd22m	102	113
#acampadalondres	111	206

Insights

Fixed Initial users = 100



- If we keep everything fixed and
  - vary the probabilities of choosing whom to mention, randomly mentioning gives the best result