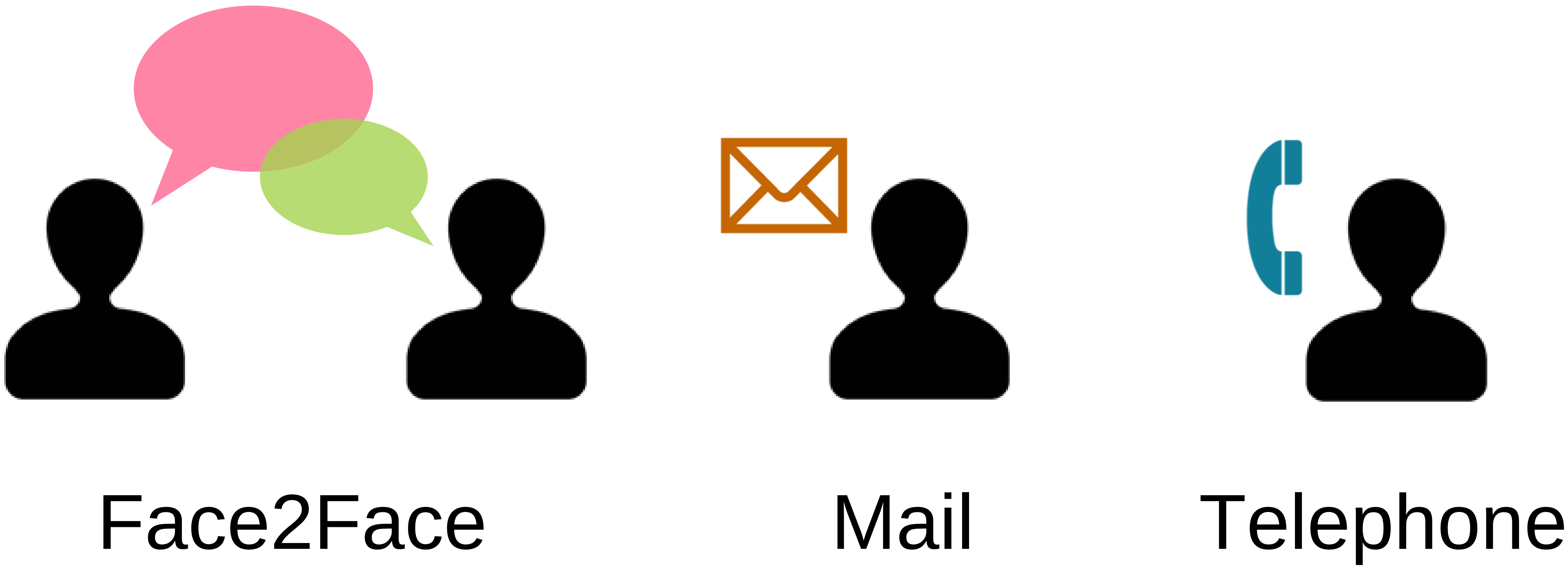
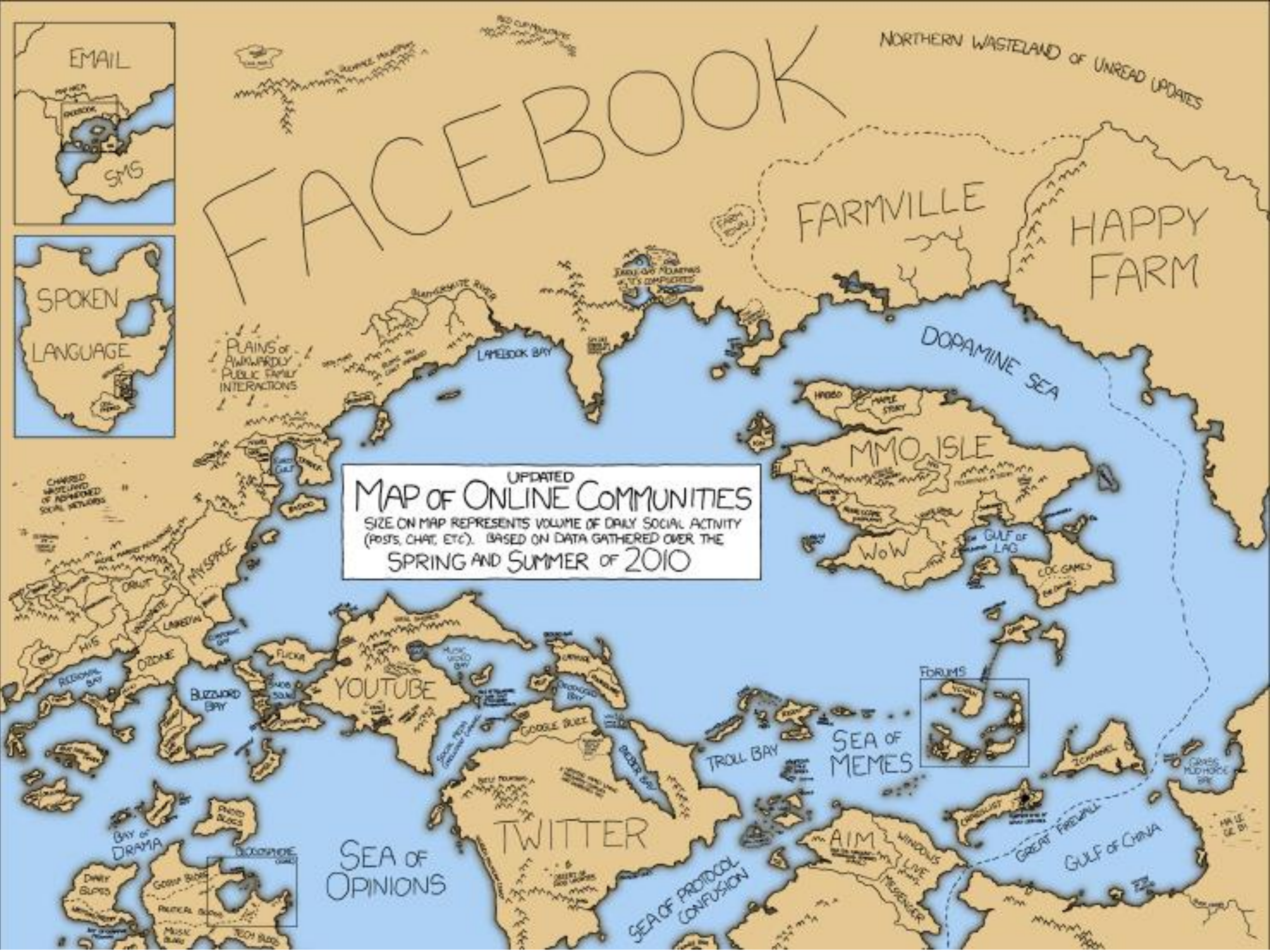
Learning Human Dynamics with Big Data from Online Social Networks

Lilian Weng
Data Scientist @ Dropbox
lilianweng.github.io

Dropbox

SCHOOL OF INFORMATICS AND COMPUTING
INDIANA UNIVERSITY
Bloomington


Face2Face
Mail
Telephone



UPDATED
MAP OF ONLINE COMMUNITIES
SIZE ON MAP REPRESENTS VOLUME OF DAILY SOCIAL ACTIVITY (POSTS, CHAT, ETC). BASED ON DATA GATHERED OVER THE
SPRING AND SUMMER OF 2010
EMAIL
SMS
SPOKEN LANGUAGE
FACEBOOK
NORTHERN WASTELAND OF UNREAD UPDATES
FARMVILLE
HAPPY FARM
DOPAMINE SEA
PLAINS OF AWKWARDLY PUBLIC FAMILY INTERACTIONS
LAMEBOOK BAY
CHARRED WASTELAND OF ABANDONED SOCIAL NETWORKS
MYSPACE
ORKUT
LINKEDIN
HI5
REGIONAL BAY
BLIZZARD BAY
FLICKR
YOUTUBE
GOOGLE BUZZ
TWITTER
MMO ISLE
HABBO
MAPLE STORY
WoW
GULF OF LAG
FORUMS
4CHAN
TROLL BAY
SEA OF MEMES
CRAIGSLIST
GREAT FIREWALL
GULF OF CHINA
AIM
WINDOWS LIVE MESSENGER
SEA OF PROTOCOL CONFUSION
SEA OF OPINIONS
BAY OF DRAMA
DIARY BLOGS
PHOTO BLOGS
BLOGOSPHERE
POLITICAL BLOGS
MUSIC BLOGS
TECH BLOGS

It becomes inexpensive and easy for people to produce, spread, and exchange information with each other.

24 PB data / Day

20 Hrs uploaded / Min

50 Mil tweets / Day

700 Bil min spent / Month

72.9 Items ordered / Sec

2.9 Mil emails / Sec

(IBM, 2012)

Computational Frameworks for Big Data

kafka

cassandra

amazon DynamoDB

APACHE GIRAPH

amazon web services™

hadoop

Google Analytics

CRM

Google bigquery

Data Warehouse

HIVE

Google Bigtable

hadoop HDFS

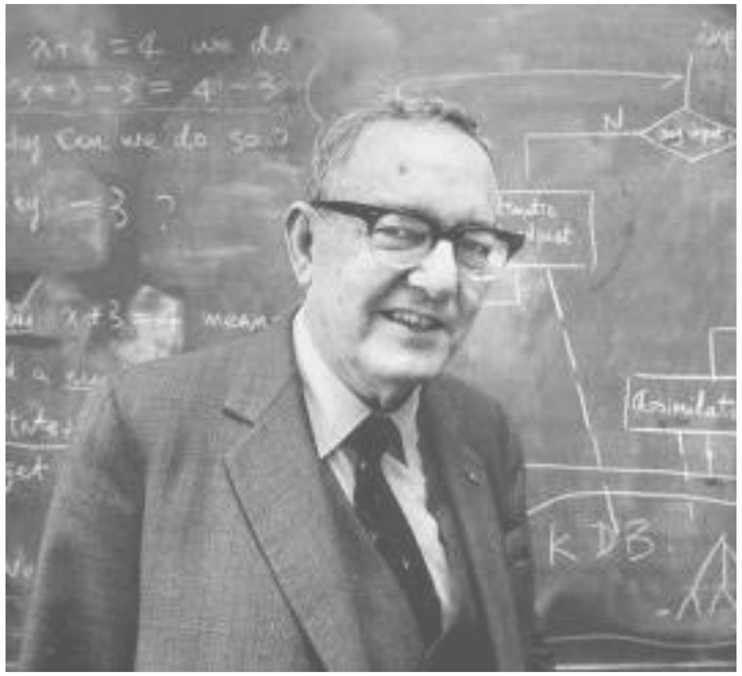
APACHE HBASE

Spark

Windows Azure


Track
Observe
Analyze
Model
Predict

Attention economy

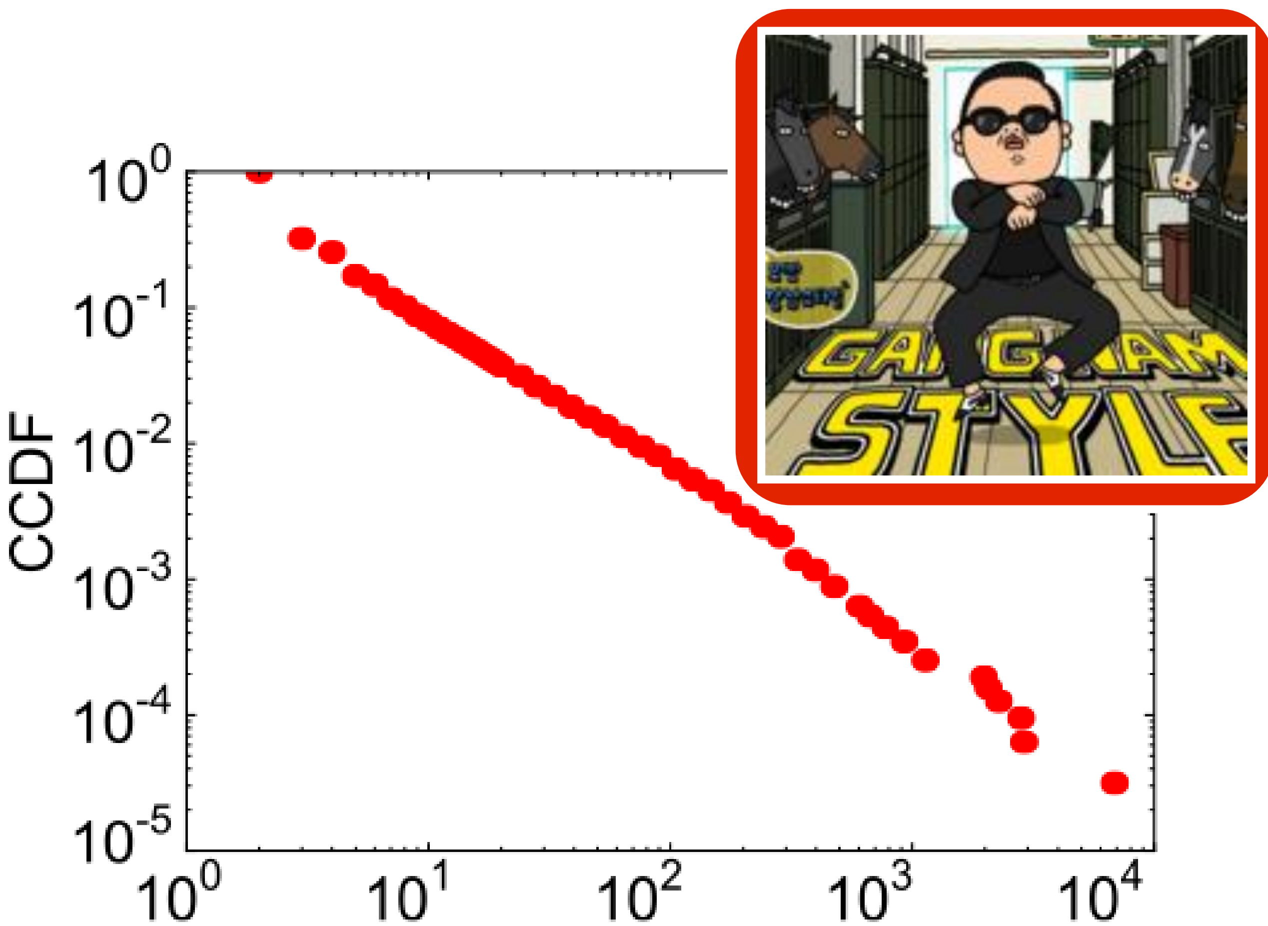


Herbert A. Simon, 1971

"What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.

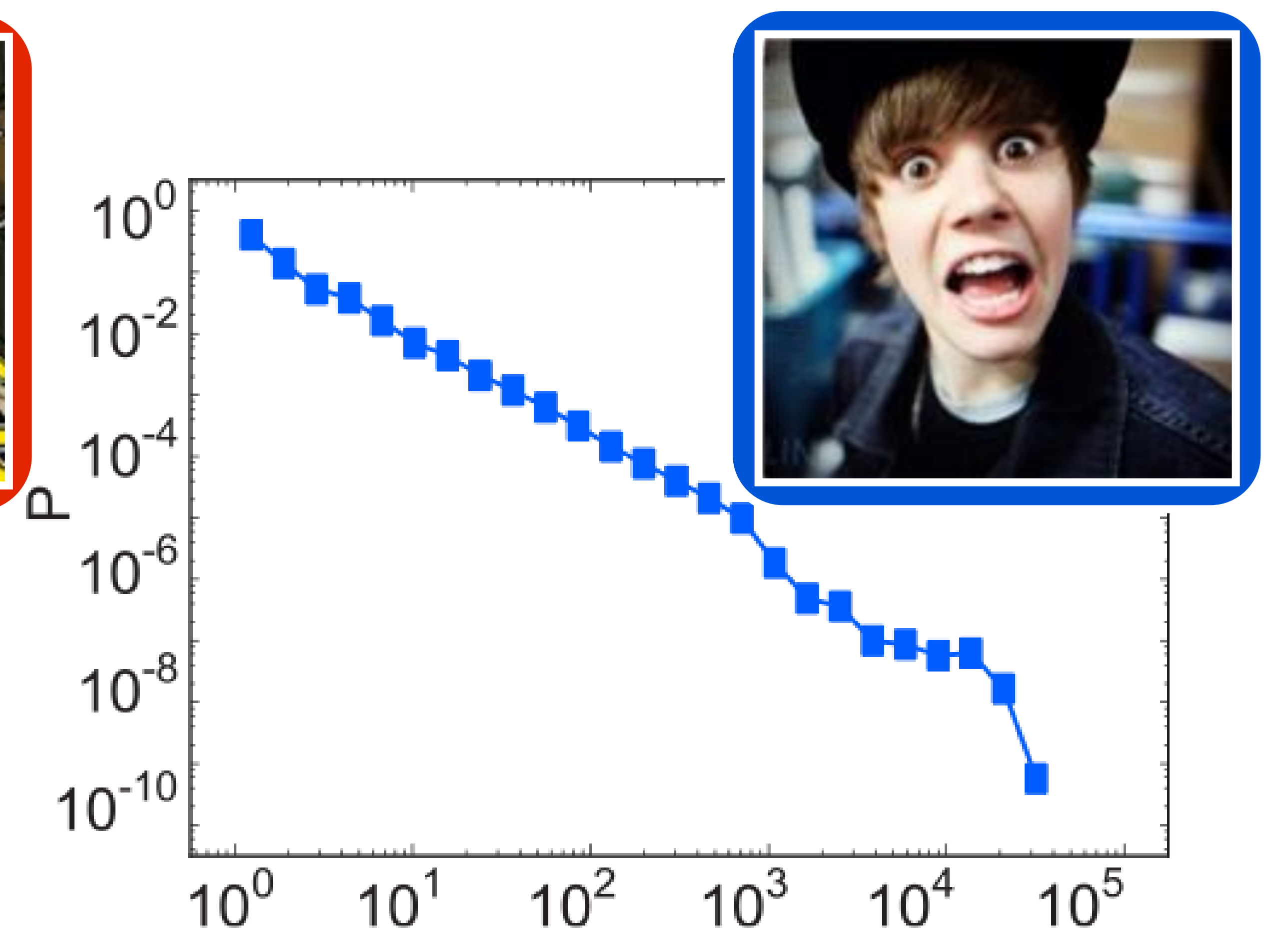
Fierce Competition but Winners still Exist

1.9 Bil Views



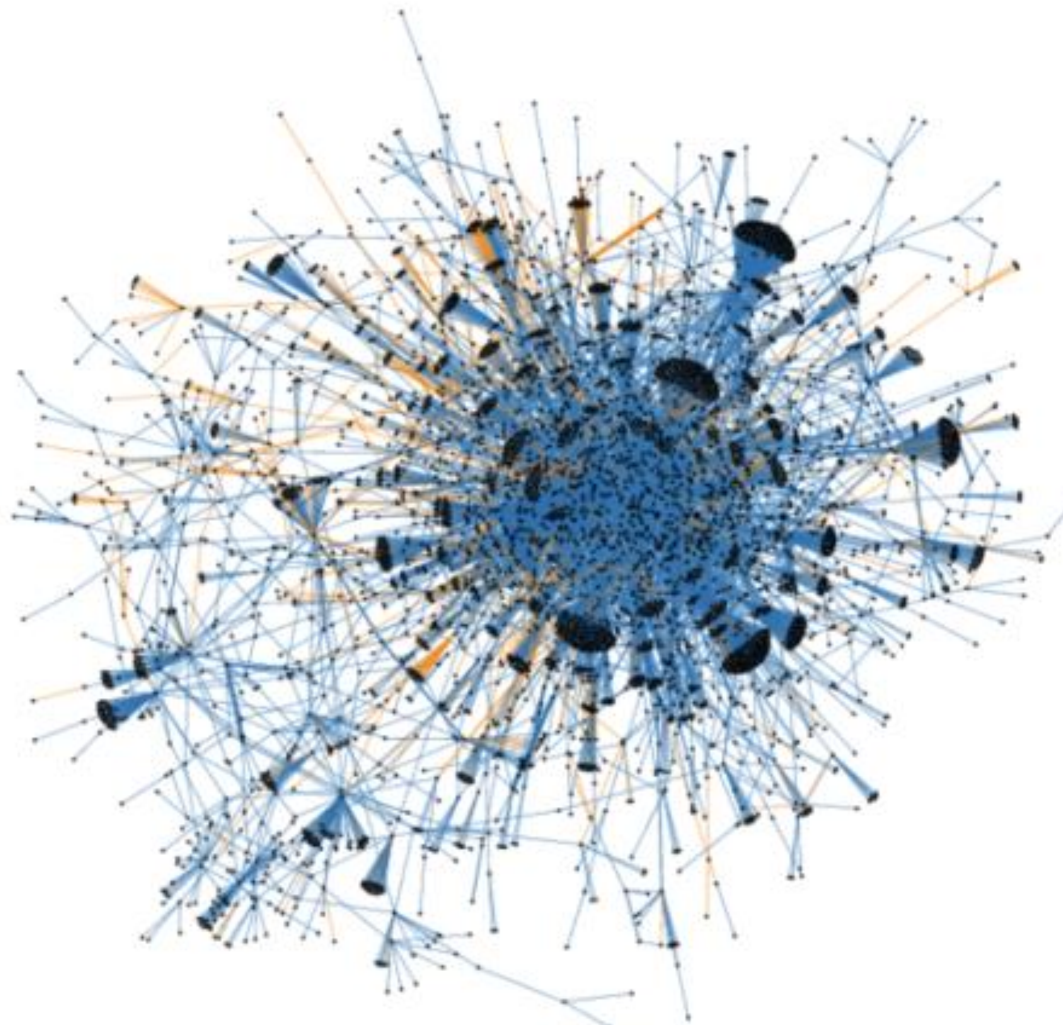
Hashtag Popularity
daily retweets
[Twitter]

50.7 Mil Followers

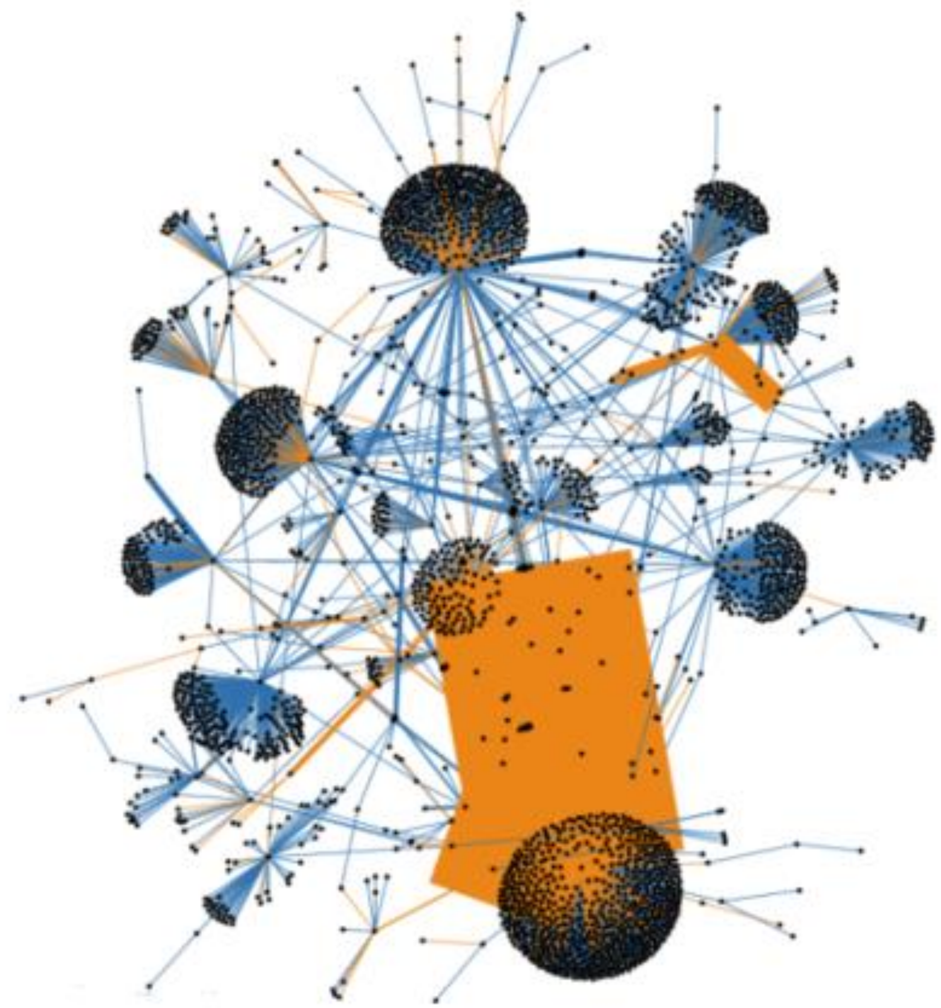


User Popularity
followers
[Yahoo! Meme]

Information diffusion happens in the wild

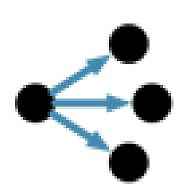


#tcot

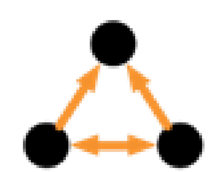


@ladygaga

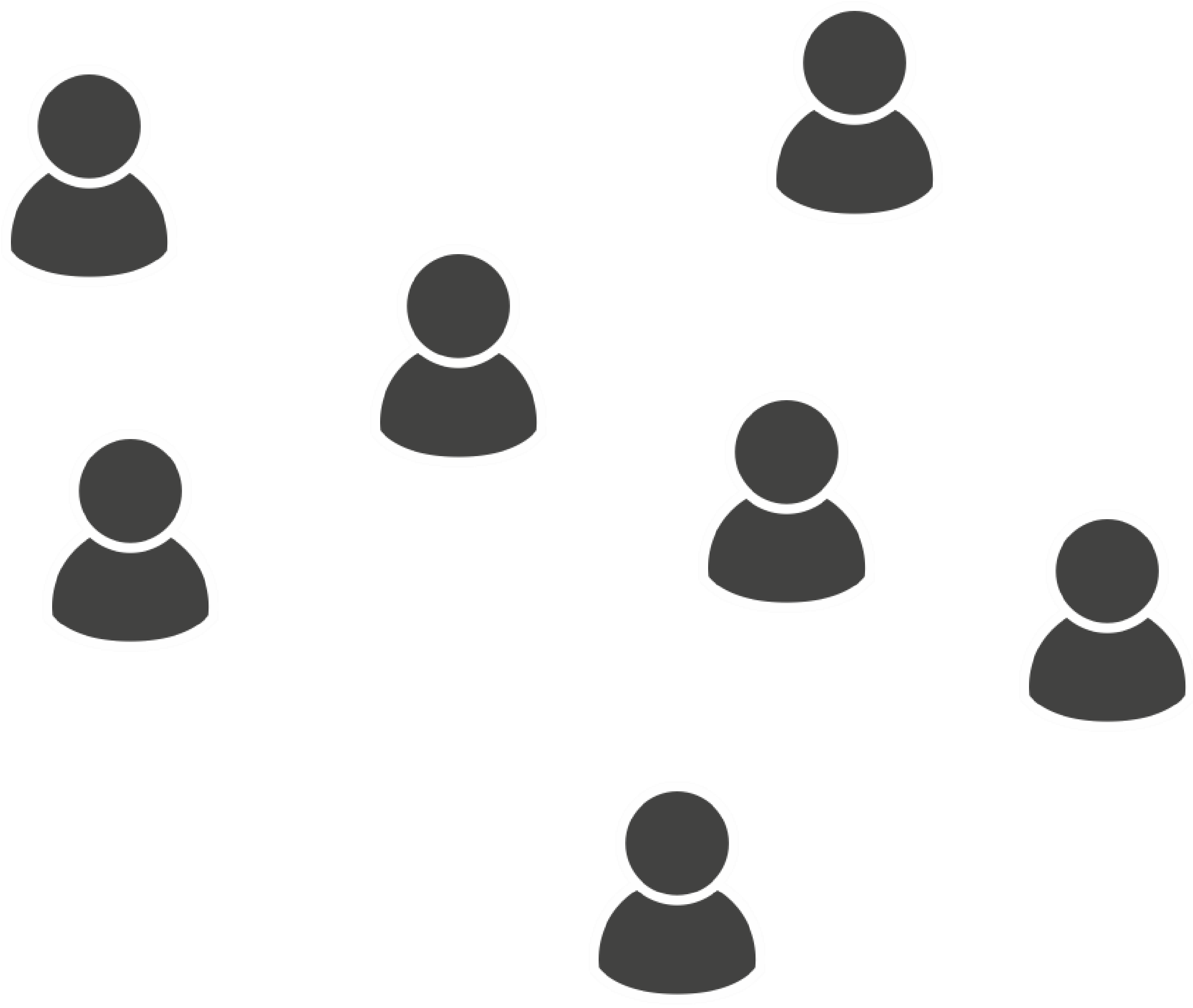
Truthy



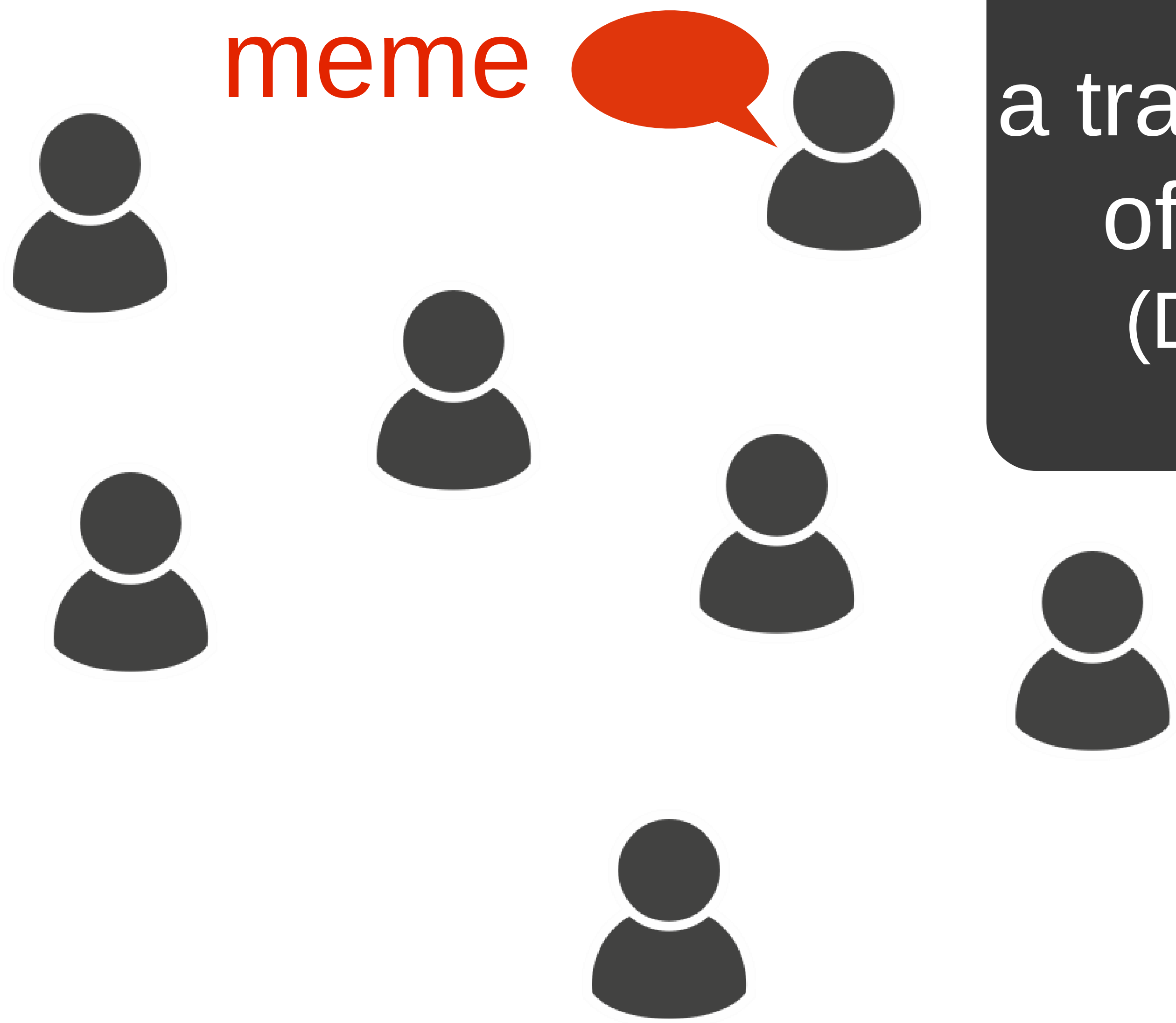
Retweet



Mention

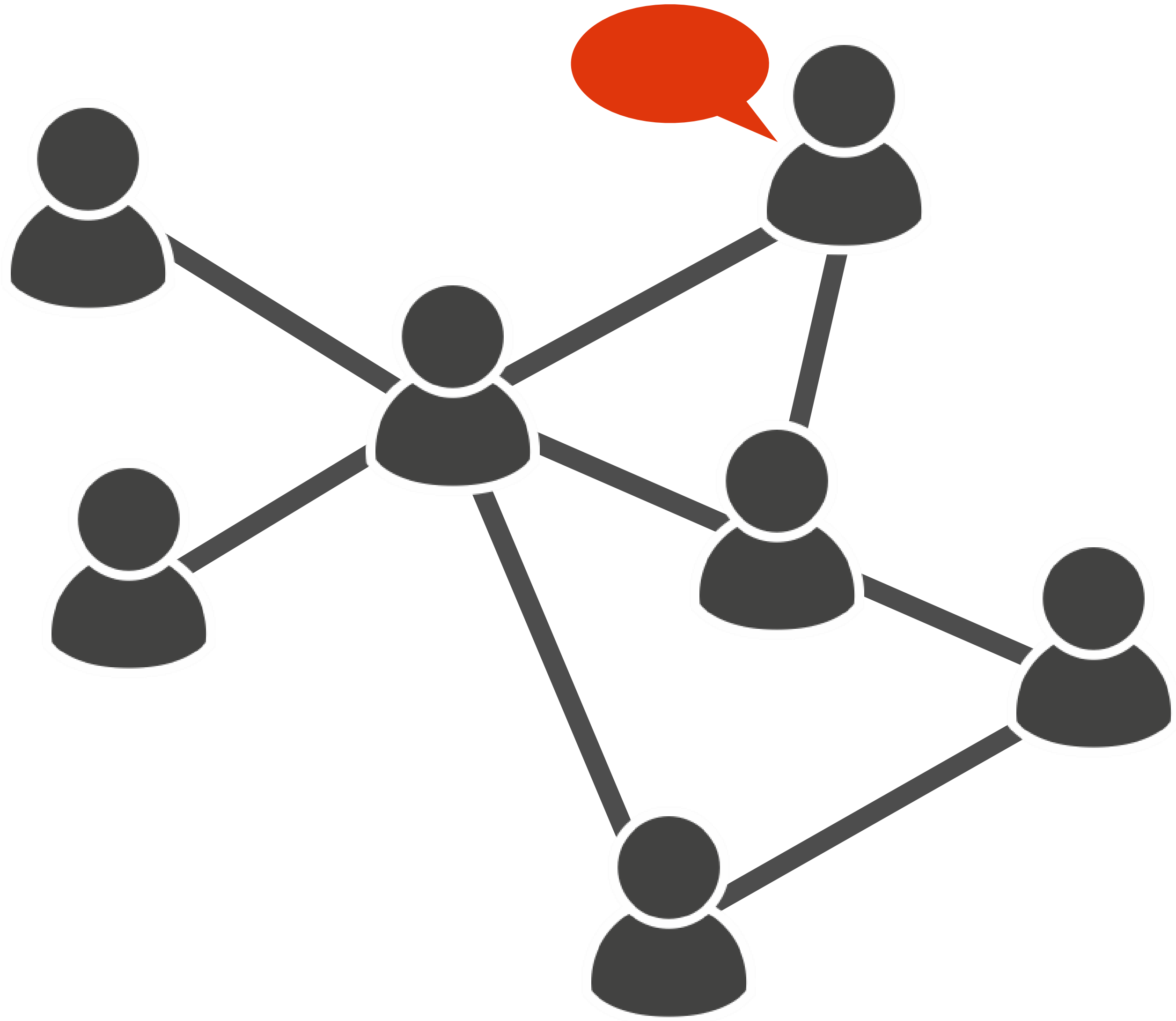


1. **People** who produce and share information

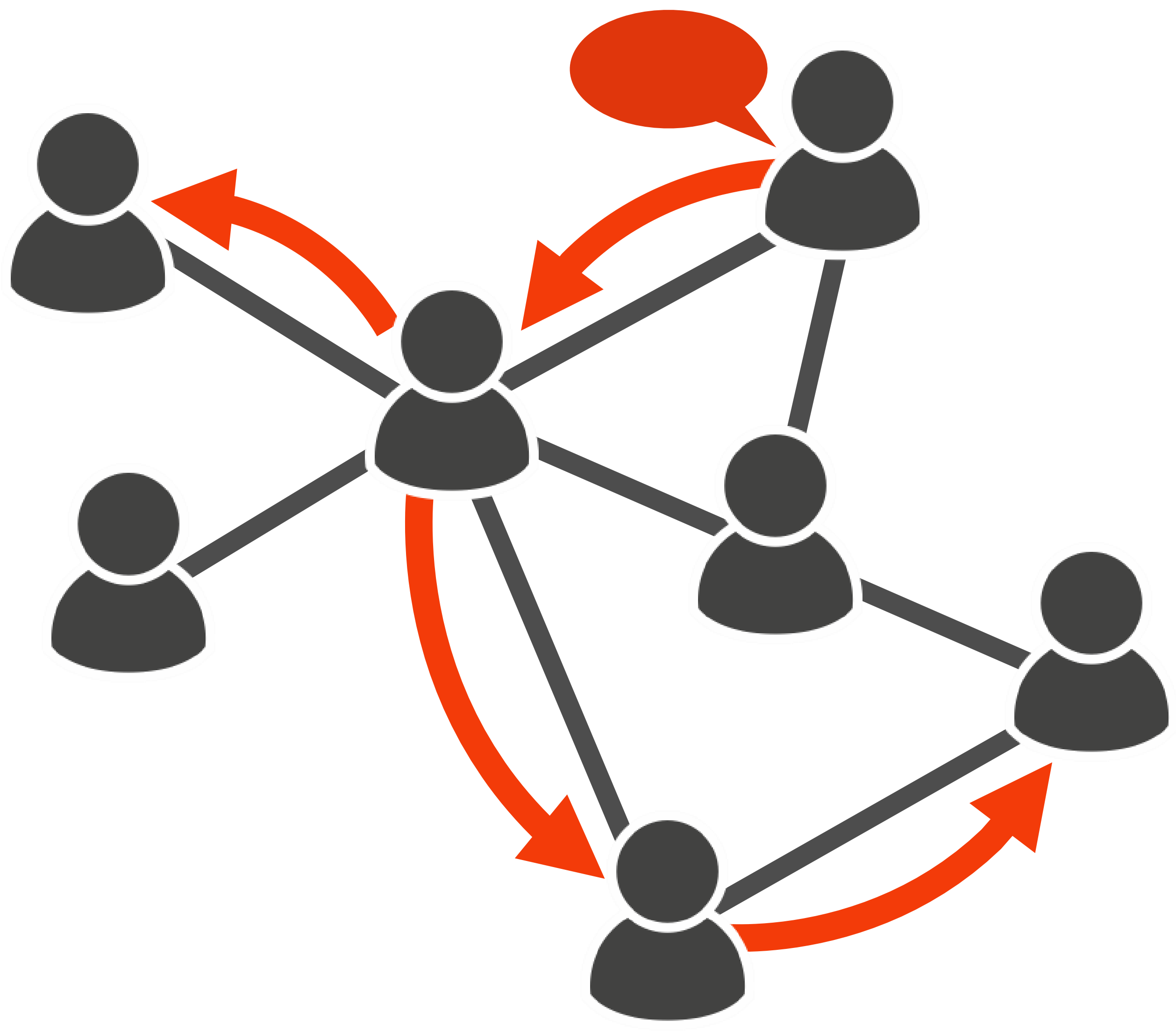


a transmissible unit of information.
(Dawkins,1989)

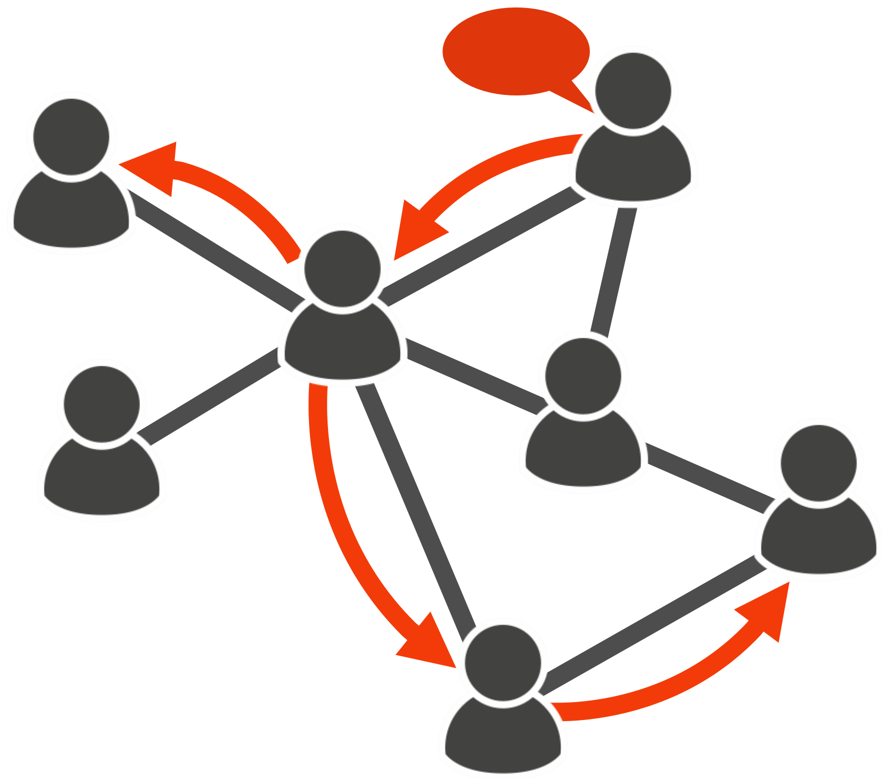
1. **People** who produce and share information
2. **Content** of transmissible messages

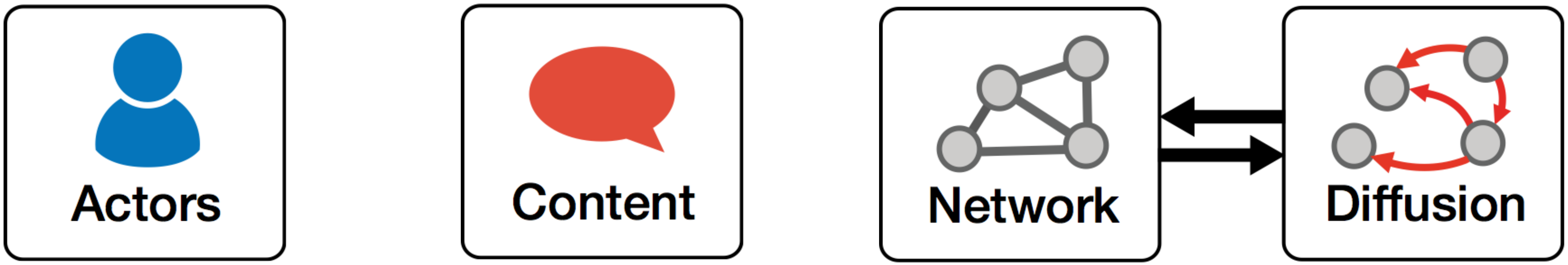


1. **People** who produce and share information
2. **Content** of transmissible messages
3. Social relationships forming the **network**



1. **People** who produce and share information
2. **Content** of transmissible messages
3. Social relationships forming the **network**
4. The mechanism of **diffusion** process



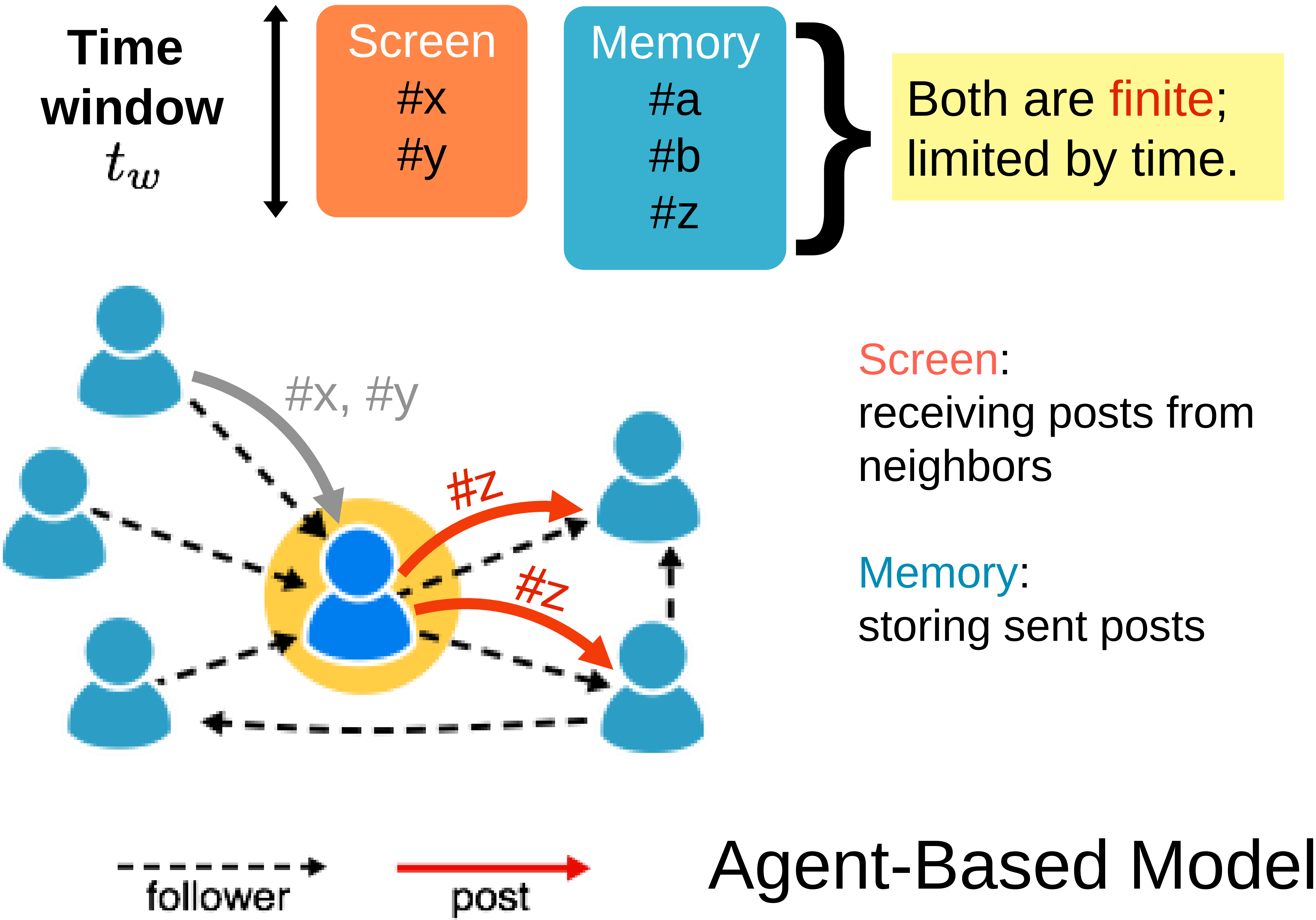


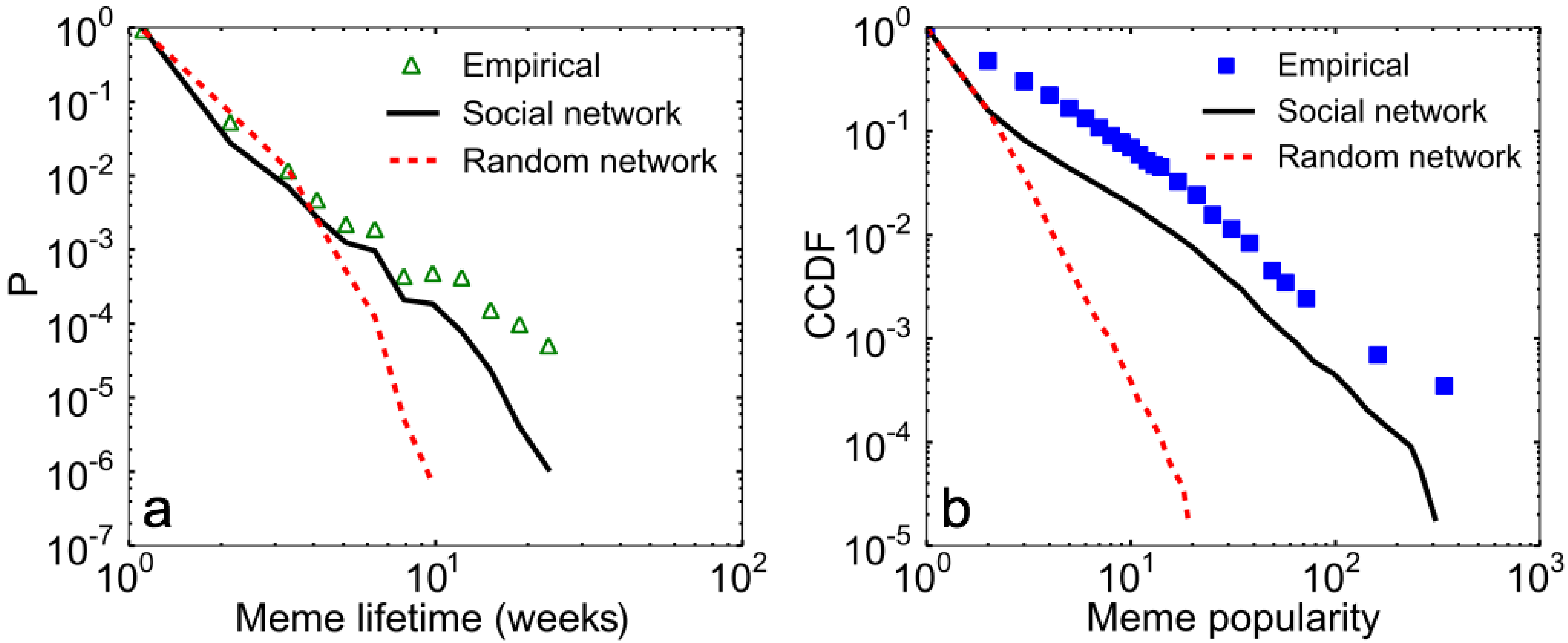
Limited attention?
Attention allocation?

Detect topics?
Topic diversity?

[1] How do network affect diffusion?
Viral meme prediction?

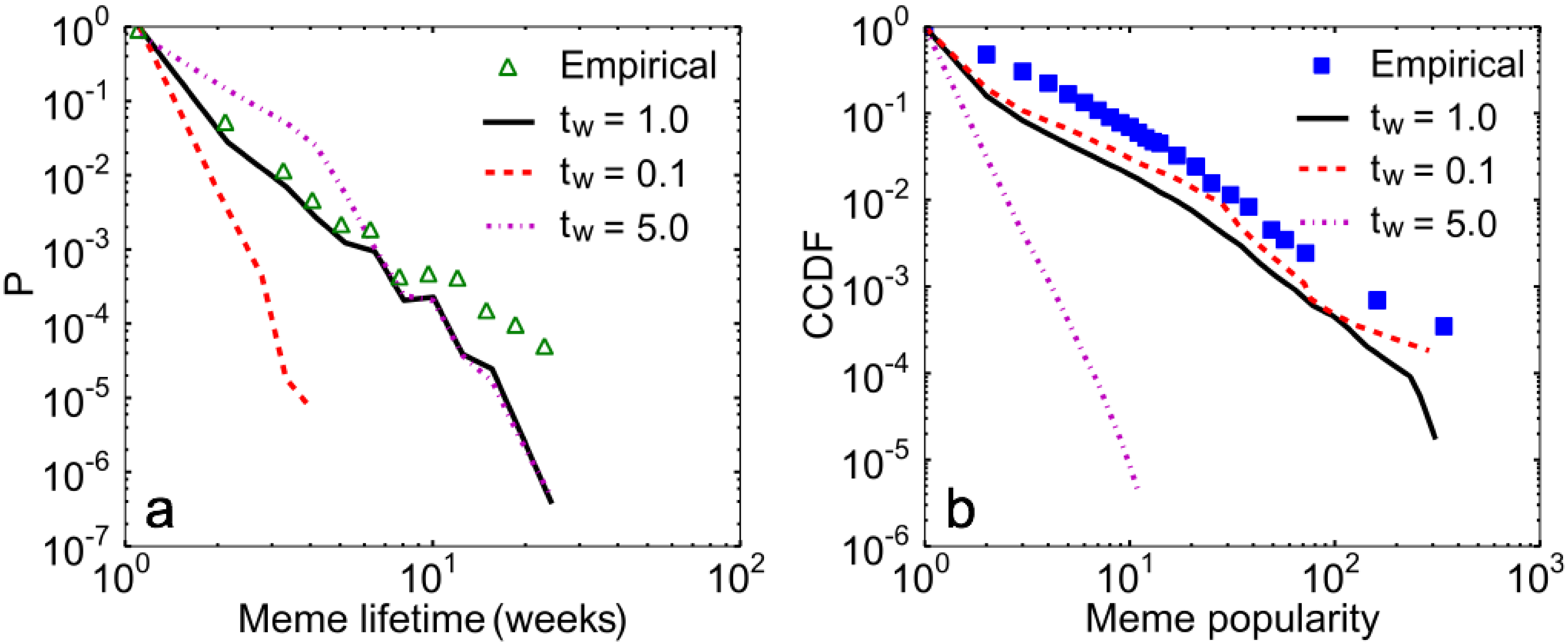
[2] How do diffusion affect network?
Traffic flows in modeling network growth?


Time window t_w
Screen
#x
#y
Memory
#a
#b
#z
Both are finite; limited by time.
#x, #y
#z
#z
Screen:
receiving posts from neighbors
Memory:
storing sent posts
follower
post
Agent-Based Model



Social network structure matters

(Weng et al. 2012)



Attention matters

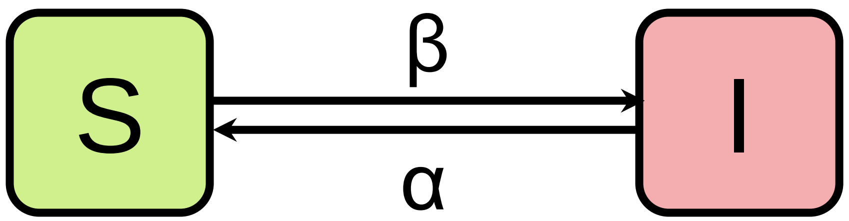
(Weng et al. 2012)

Social network structure + Competition for limited attention

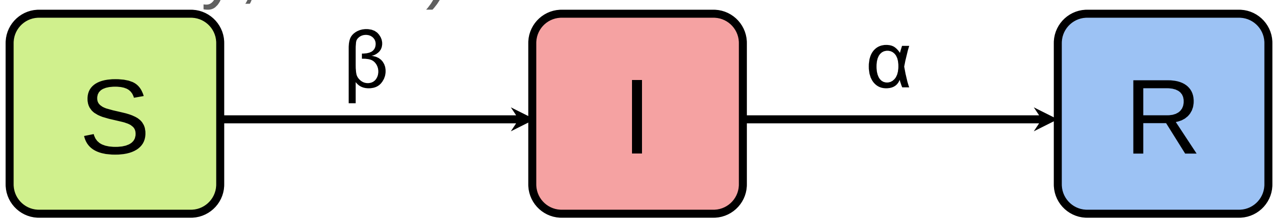
Heterogeneity of meme dynamics

Information Diffusion

- The SIS Model (Bailey,1975)



- The SIR Model (Anderson & May,1992)

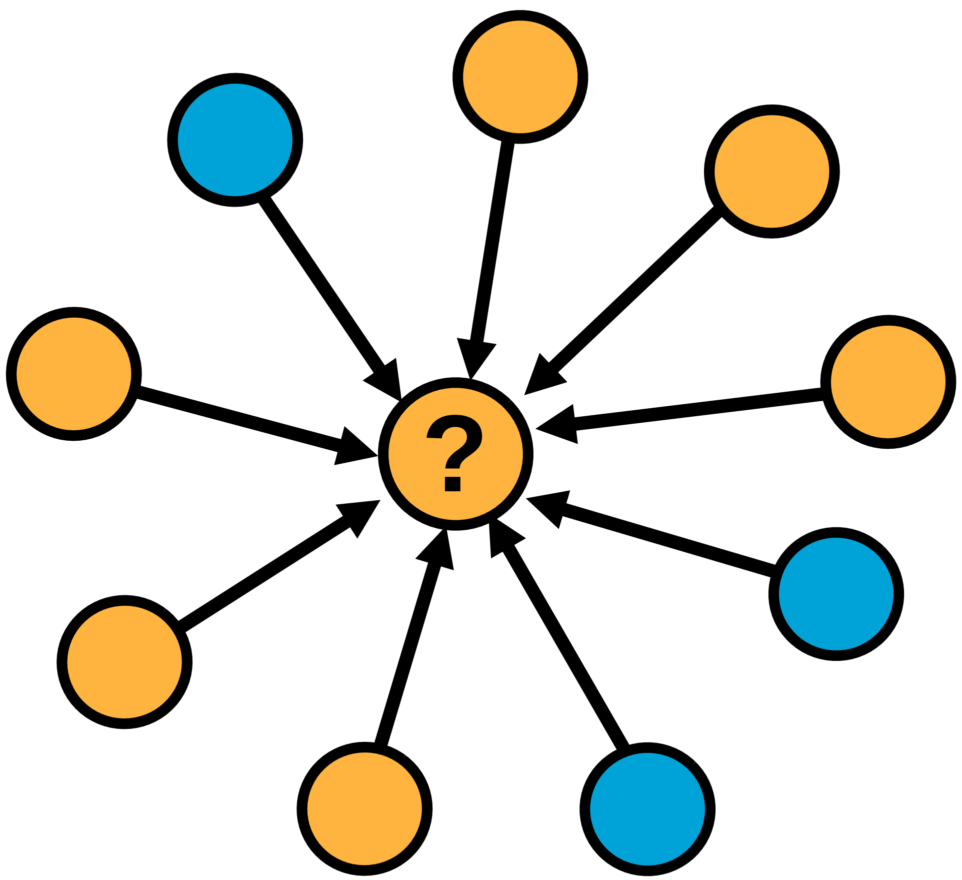


Epidemic models

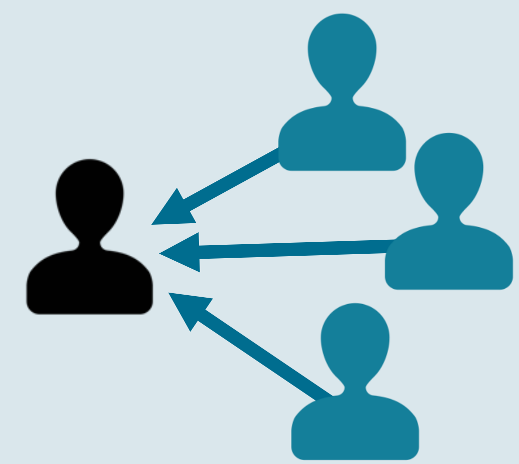
Diseases

Simple contagion

Information Diffusion



Threshold model
(Granovetter, 1978)

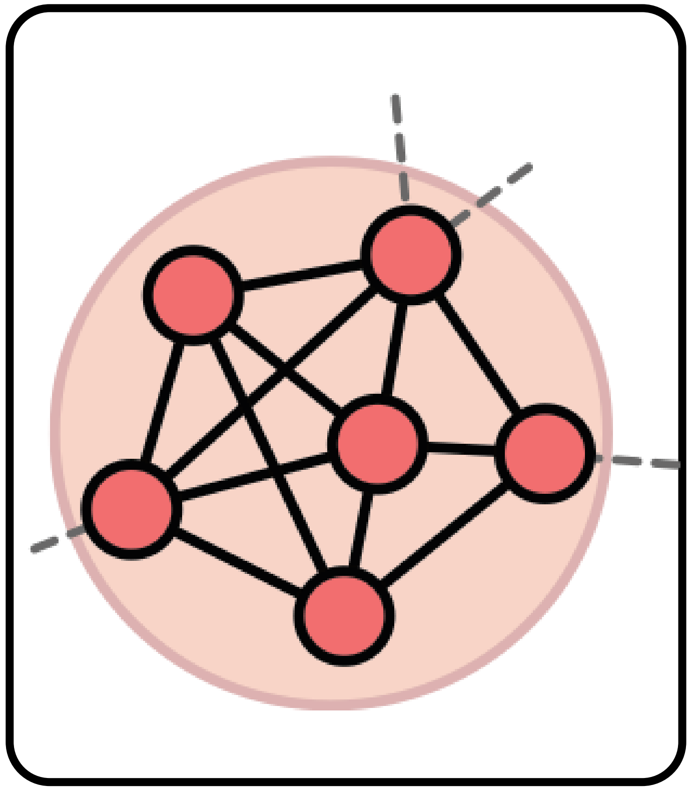


Ideas or behavior:
Complex contagion

- DBLP (Backstrom et al., 2006)
- Twitter (Huberman et al., 2008; Romero et al., 2011)
- Wikipedia (Cosley et al., 2010)
- Facebook (Ugander et al., 2012)

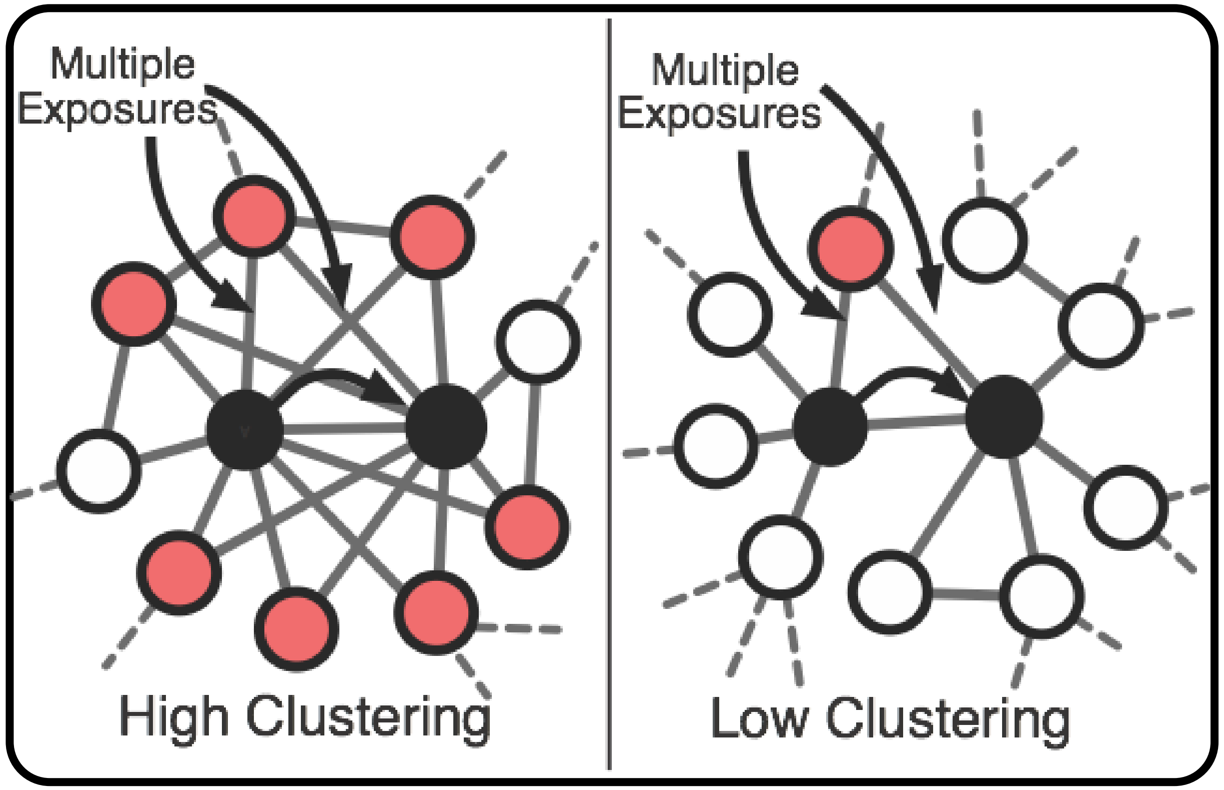
Community Trapping Effect

Structural Trapping



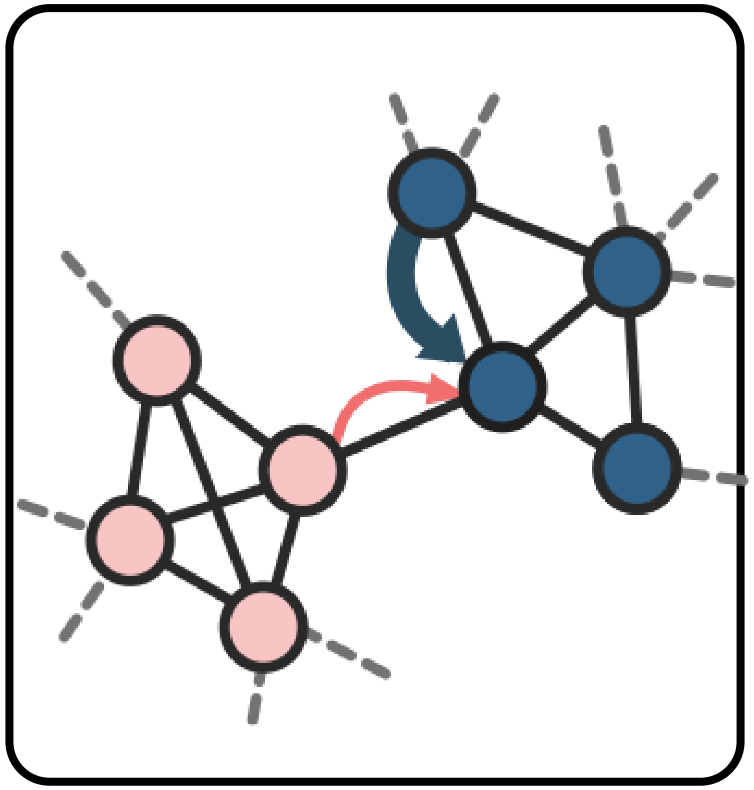
Social Reinforcement

(Centola, 2010)



Homophily

(McPherson et al.,2001)



Null Models

	Community trapping effects		
	Network	Reinforcement	Homophily
M1: Random distribution			
M2: Random diffusion	√		
M3: Social reinforcement	√	√	
M4: Homophily	√		√

Null Models

	Community trapping effects		
	Network	Reinforcement	Homophily
M1: Random distribution			
M2: Random diffusion	√		Simple contagion
M3: Social reinforcement	√	√	
M4: Homophily	√		√

Null Models

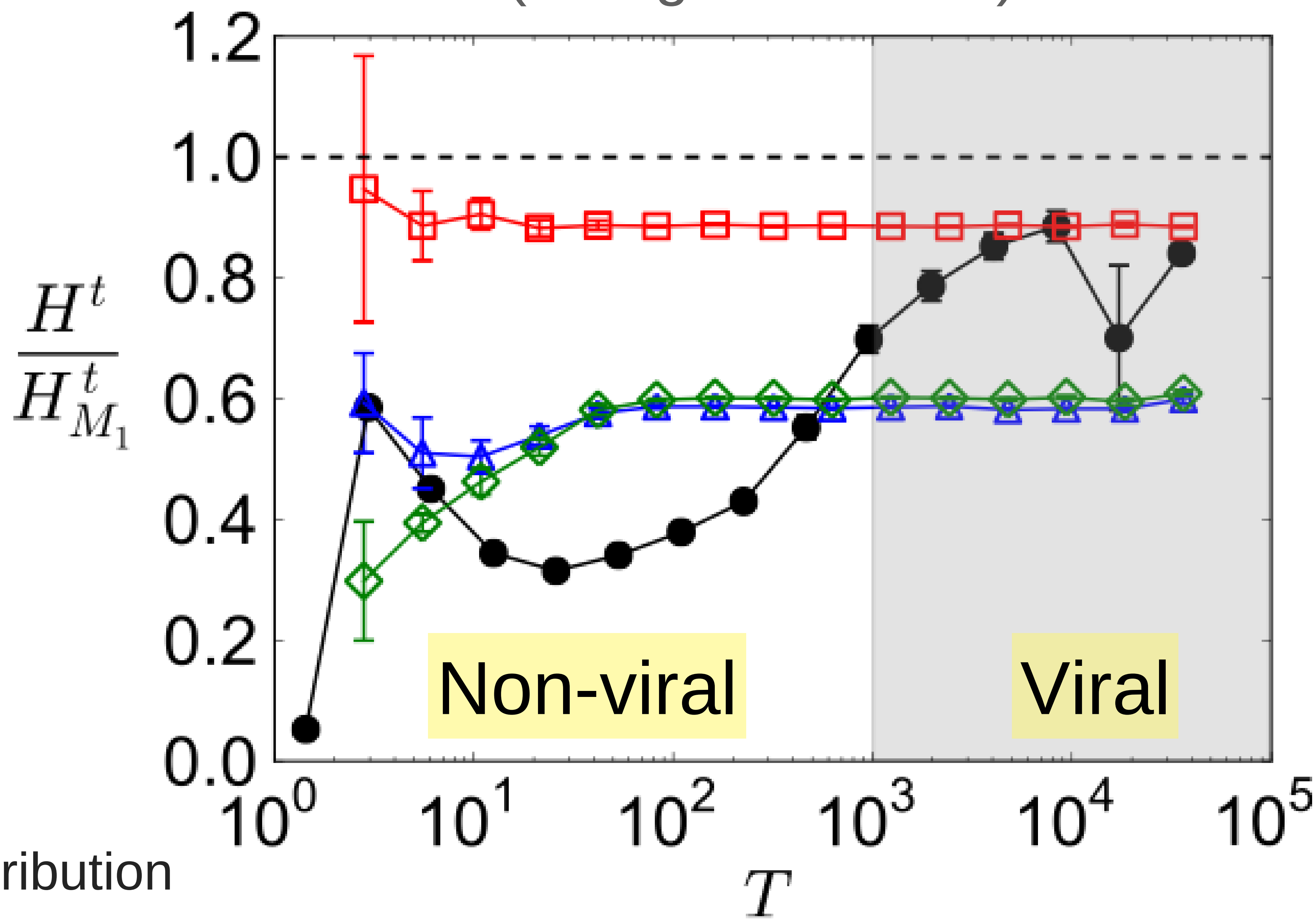
	Community trapping effects		
	Network	Reinforcement	Homophily
M1: Random distribution			
M2: Random diffusion	√		
M3: Social reinforcement	√	√	
M4: Homophily	√		√

Complex contagion

Relative Usage Entropy

(Weng et al. 2013)

Entropy of # tweets distributed in different communities



M1: Random distribution

M2: Random diffusion

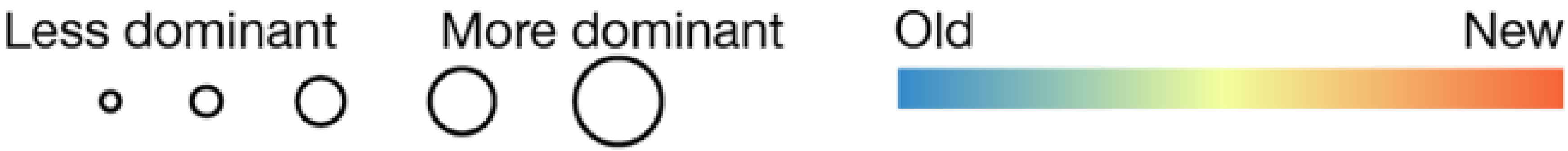
M3: Social reinforcement

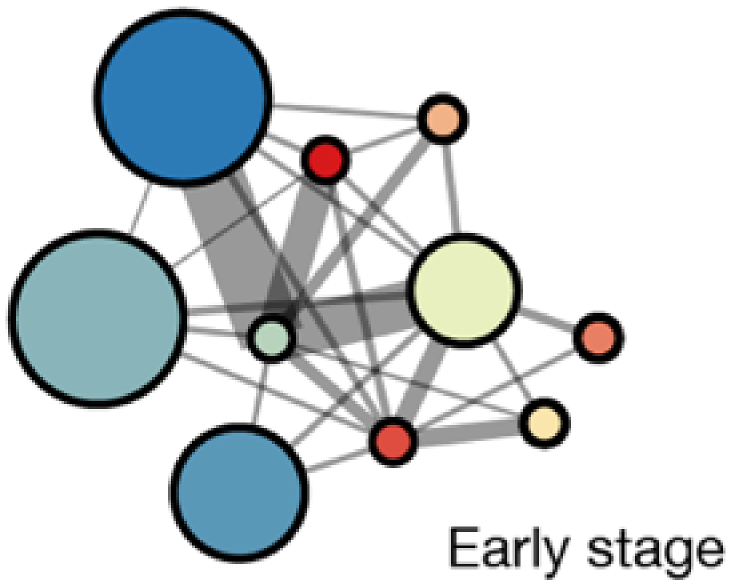
M4: Homophily

Total # tweets

Viral memes are less trapped by communities, more like disease.

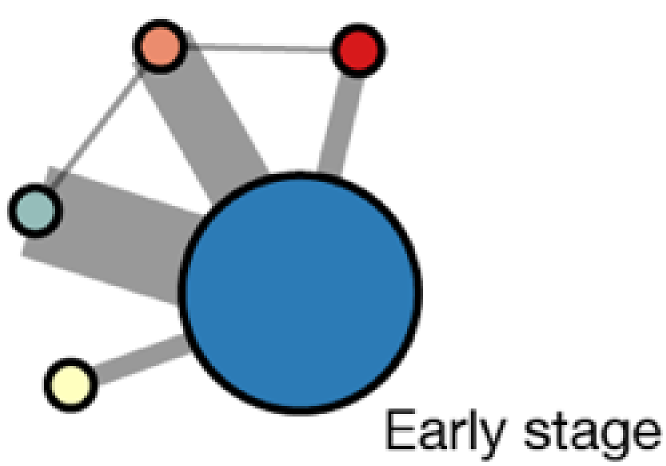
Can we predict the future meme virality by qualifying concentration across communities?





#ThoughtsDuringSchool

30 tweets



#ProperBand

30 tweets

Virality Prediction

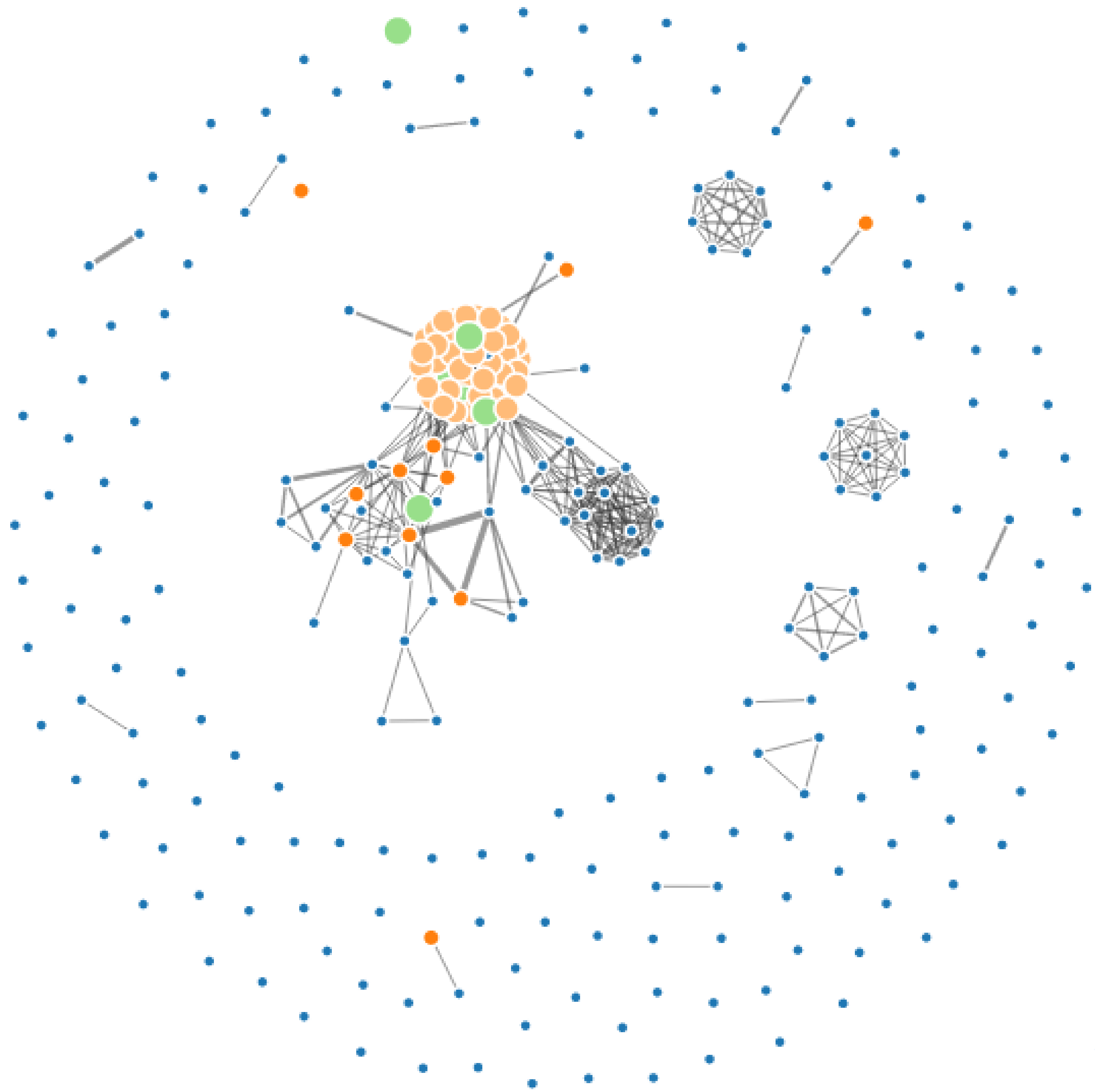
1 Community-blind features
- # Early adopters
- Size of infection frontier

Binary classification
Predict whether a meme is viral (>1000 tweets)

2 Community-based features
- # Infected communities
- Entropy
- Frac. intra-community RT/@

(Weng et al. 2013)

Collaboration Network @ Dropbox

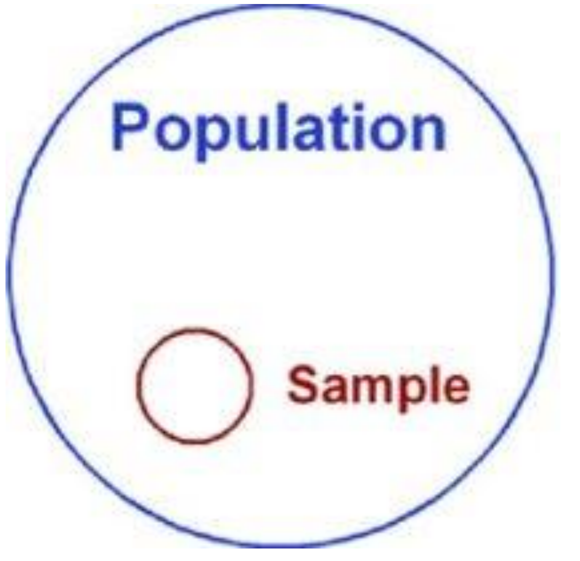


Dropbox

Big Data Challenges

- Data Sampling
- Universality
- Privacy
- Open Access
- Gap between Online and Offline Systems

Data Sampling



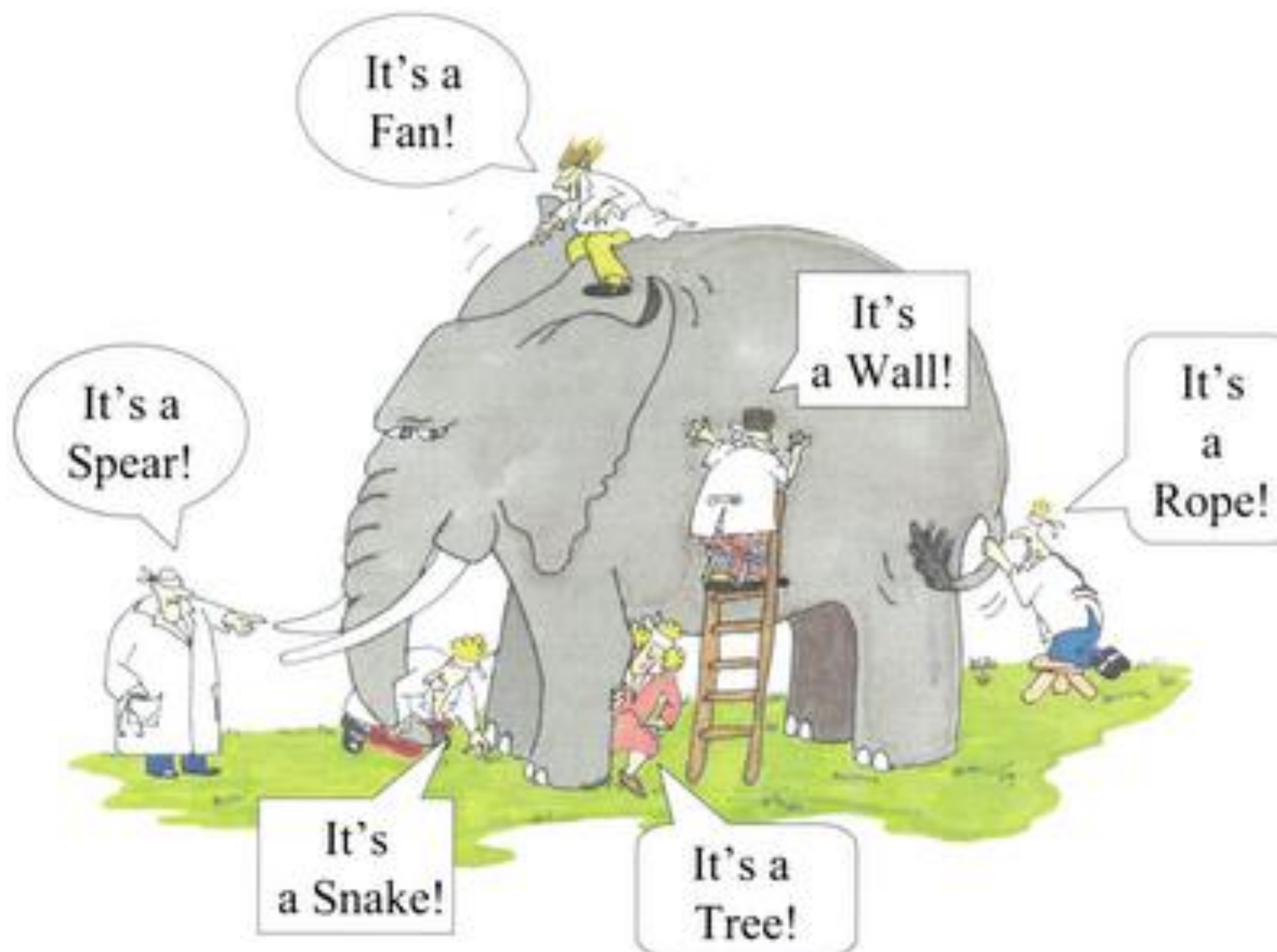
- Most studies involve sampled datasets.
- Good or poor representation of the system?
- Incorrect sampling could lead to biased results.

Universality

- Most studies only used a single system or a snapshot of the system.
- "*blind men feeling the parts of an elephant*" (Lazer et al., 2009)

Universality

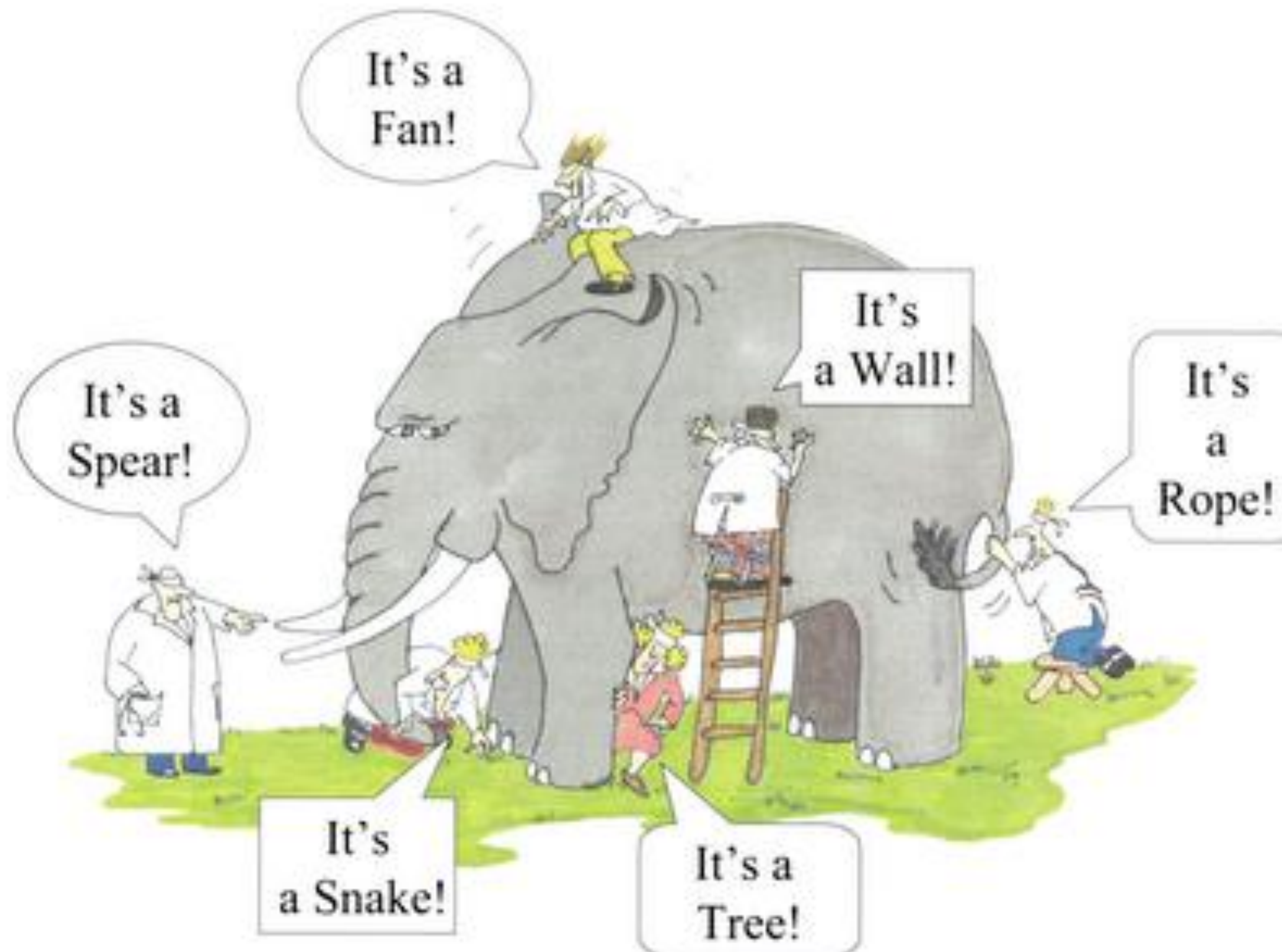
- Most studies only used a single system or a snapshot of the system.



Universality

- Most studies only used a single system or a snapshot of the system.

- More future work is expected to study the longitudinal patterns on data with long history and to compare multiple platforms.



Privacy

- People exposure more personal information online.
- Look across data from multiple sources to decipher the trace of an individual user.
- Occupation, address, birth date, and social security number, personal schedules

Open Access

- Data is crucial in quantitative research.
- Some datasets cannot be public.
- No external replication or verification of the findings.
- Balance between open environment and privacy concerns.

Gap between Online and Offline Systems

- Online behavior is usually well curated and systematically managed [Ellison et al., 2006].
- Can we safely apply classical sociological theorems to online systems, or extend the findings derived from online big data to offline social movements and events?

Selected Papers

- L. Weng, A. Flammini, A. Vespignani, & F. Menczer. Competitions among topics in a world with limited attention. Nature **Sci. Rep.**, (2)335, 2012.
- L. Weng, et al. The Role of Information Diffusion in the Evolution of Social Networks. In: **KDD**. 2013.
- L. Weng, F. Menczer, & Y.-Y. Ahn. Virality Prediction and Community Structure in Social Networks. Nature **Sci. Rep.**, (3)2522, 2013.
- L. Weng, F. Menczer, & Y.-Y. Ahn. Predicting Meme Virality in Social Networks using Network and Community Structure. In: **ICWSM**. 2014.
- L. Weng & T. Lento. Topic-based Clusters in Egocentric Networks on Facebook. In: **ICWSM**. 2014.
- L. Weng & F. Menczer. Topicality and Social Impact: Diverse Messages but Focused Messengers. **PLOS ONE**. 2015.

Check http://lilianweng.github.io for more and details.

Thank You!
Questions?

Sincerely, Lilian