# Virality Prediction and Community Structure in Social Networks


Yong-Yeol "YY" Ahn
@yy

SCHOOL OF INFORMATICS AND COMPUTING
INDIANA UNIVERSITY
Bloomington


LinkedIn
ursquare
LIVEJOURNA
vimeo
Tube
Insta
facebook
Blogger
my
flickr
Bēha
tumblr.
Like

# Most populated countries

1,300,000,000+

1,200,000,000+

300,000,000+

# Most populated countries

| | |
|---|---|
|  | 1,300,000,000+ |
|  | 1,200,000,000+ |
| facebook | 1,100,000,000+ |
| | 500,000,000+ |
|  | 300,000,000+ |

Social media can **amplify** messages.

“Casually pepper spraying everything cop” meme


The Constitution
We the People
WALL


“This Revolution Will Be Tweeted.”


Facebook


WHO'S
AFRAID OF
TWITTER?!

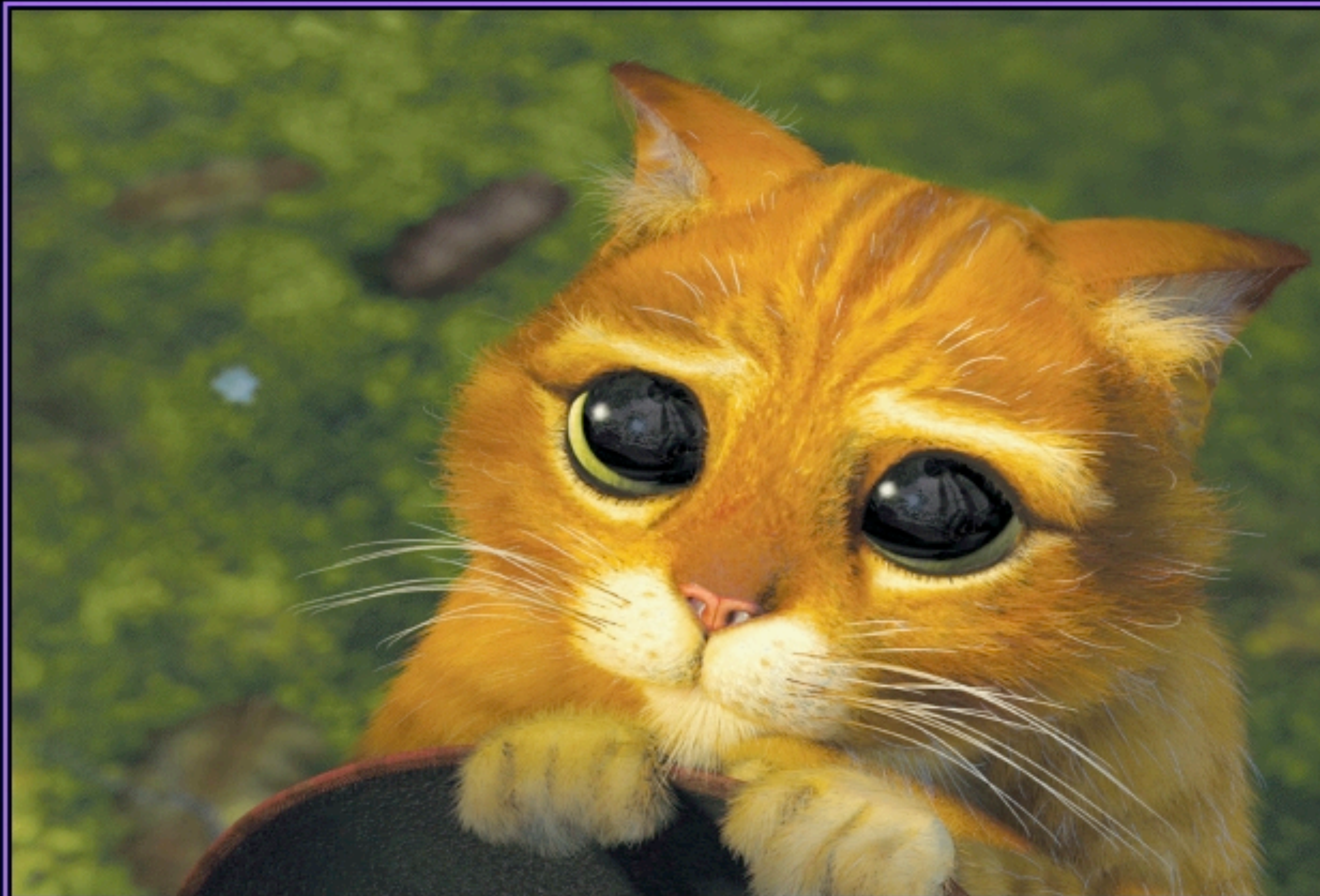

Is Oprah's Network Too White? ▪ The Mammogram Hustle
FEBRUARY 7, 2011
Newsweek
Rage Goes Viral
From Tunisia to Egypt to Yemen, a youthquake is rocking the Arab world. Get ready for the aftershocks.
Will Business Buy What Obama's Selling?
The Dirty Secret of Apple's Design
America's Happiest Unconventional Family

Companies are desperately trying to leverage **social media**

to make their products and ads ***viral.***

Original, useful ideas
- hard

“viral marketing”



VOTE FOR ME.!

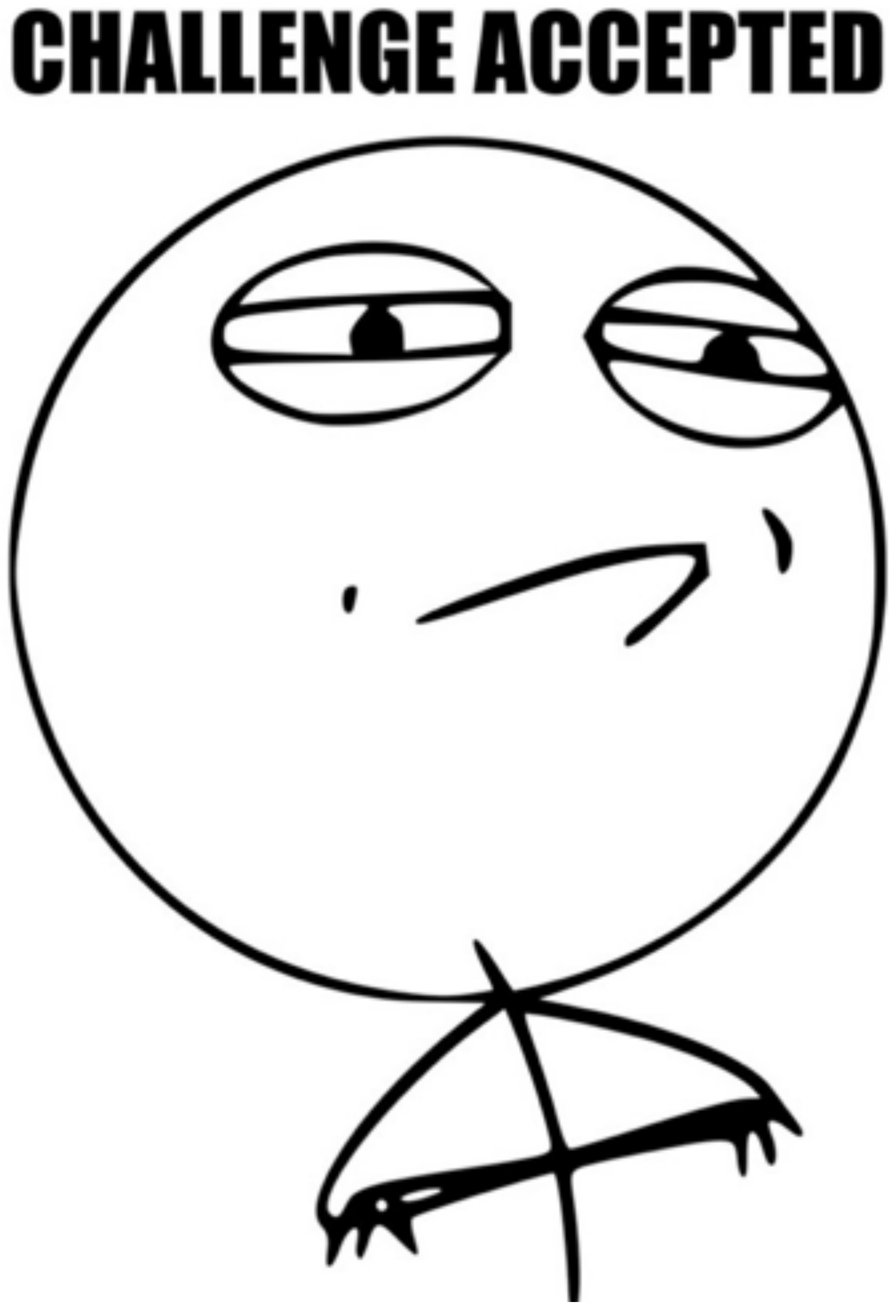
‘Astroturfing’ may change election results.

“1/3 of online reviews may be fake.” - Bing Liu (UIC)

# KEEP
# CALM
AND
# GANGNAM
# STYLE

How can we understand ***virality***?

Can we **predict** viral memes?


CHALLENGE ACCEPTED

Clue #1:

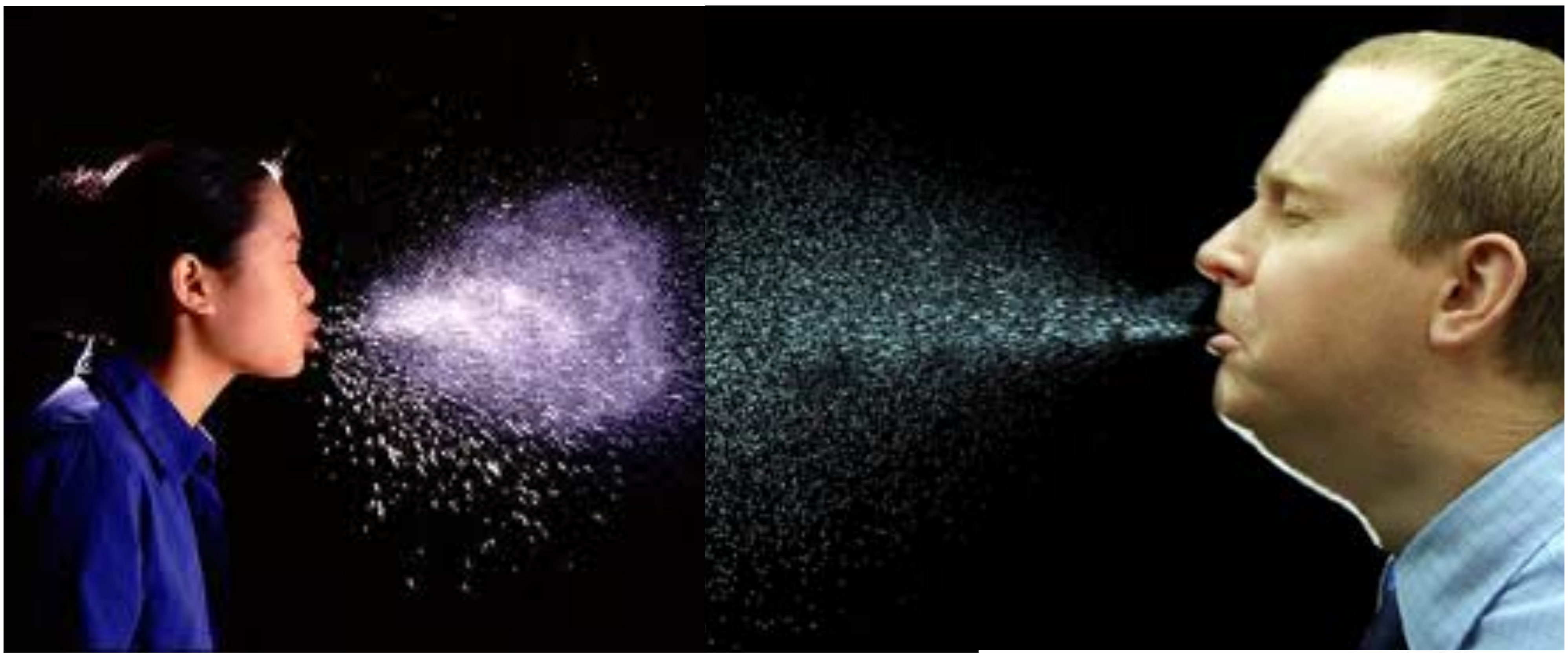
**Complex contagion**

Memes

=

Infectious diseases?

# Germs spread through the "**social**" network



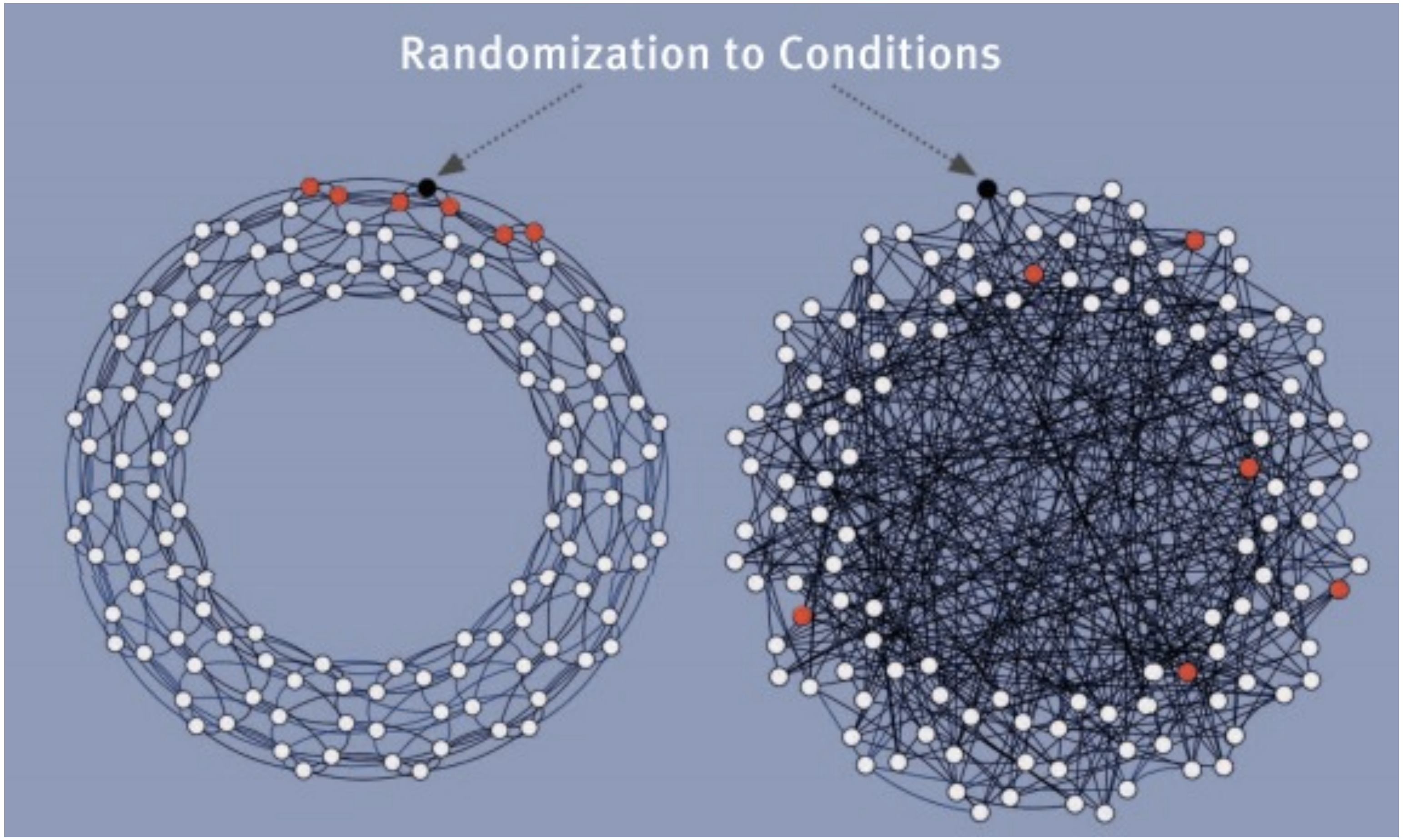
Memes, ideas and behaviors also **spread** through **social network**.

Memes

=

Infectious diseases?

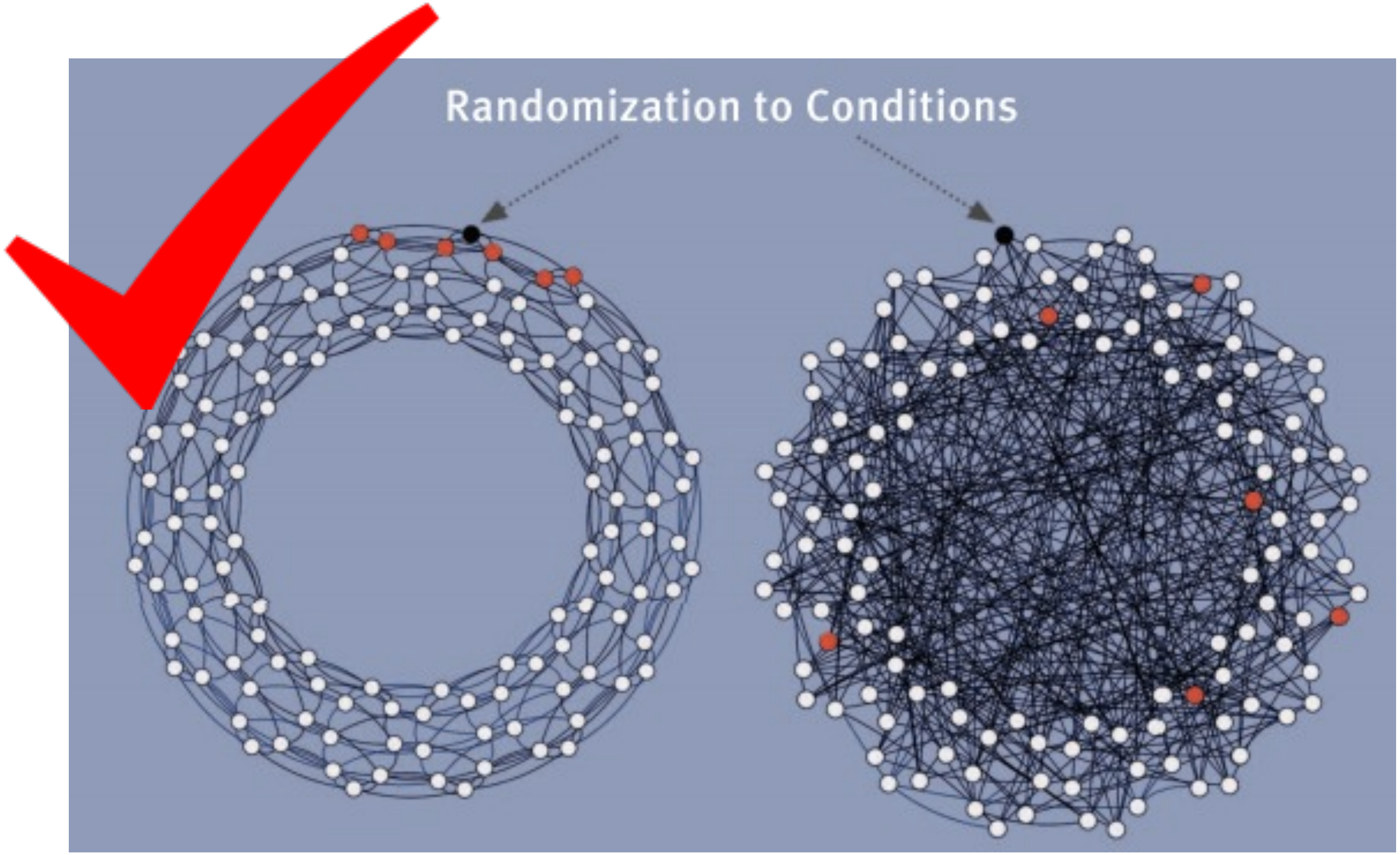
Maybe not.




“Large” world “Small” world

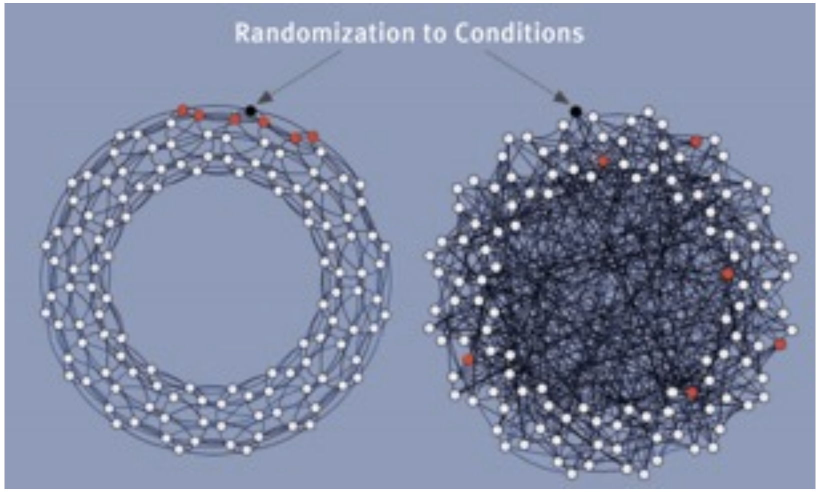
D. Centola, *Science* 2010

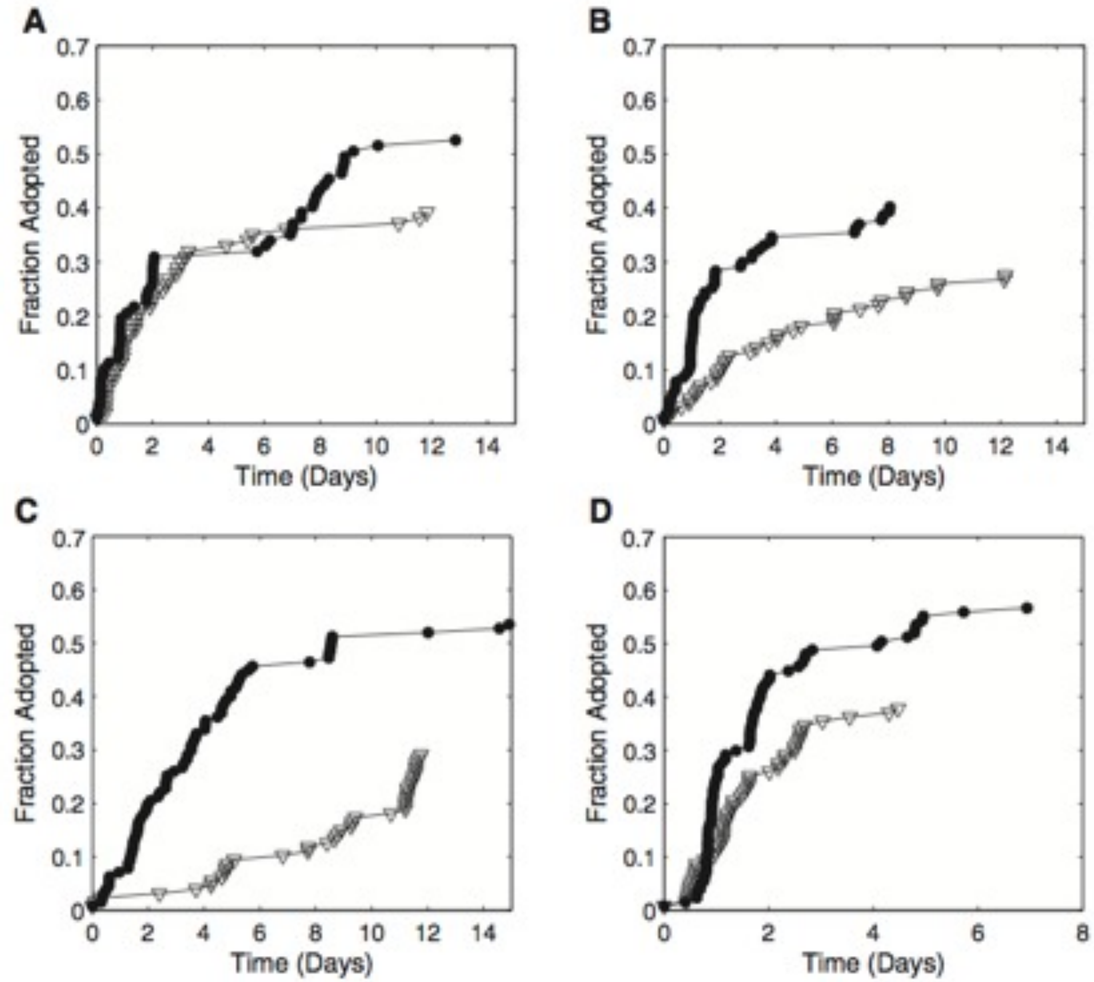
Which network is better at spreading information quickly?

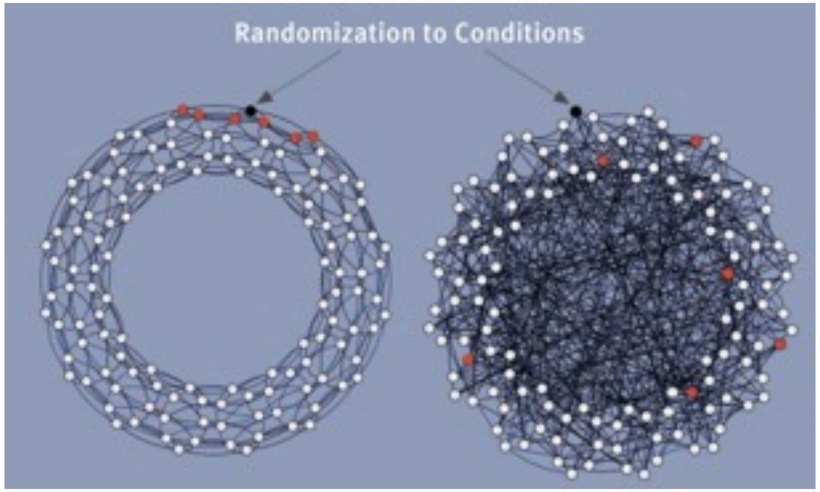


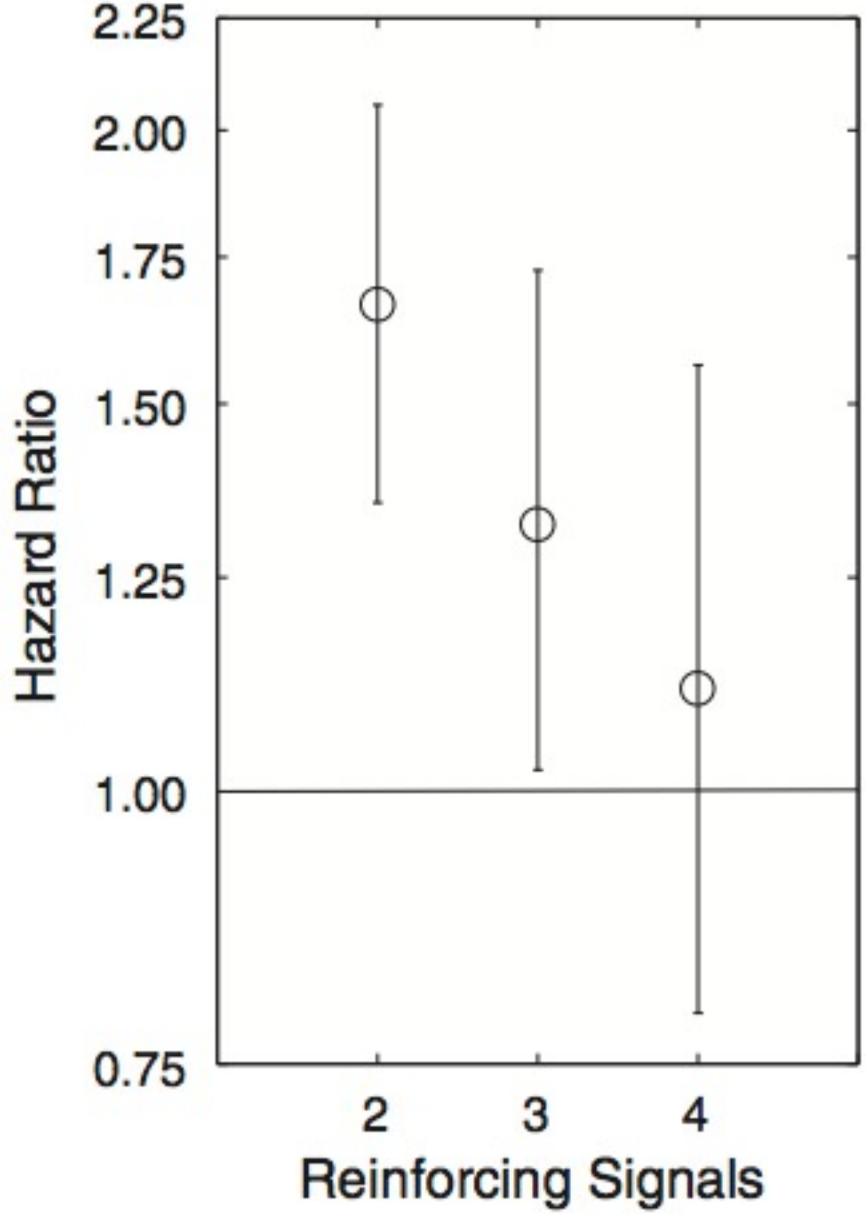

“Large” world “Small” world

D. Centola, *Science* 2010


Randomization to Conditions

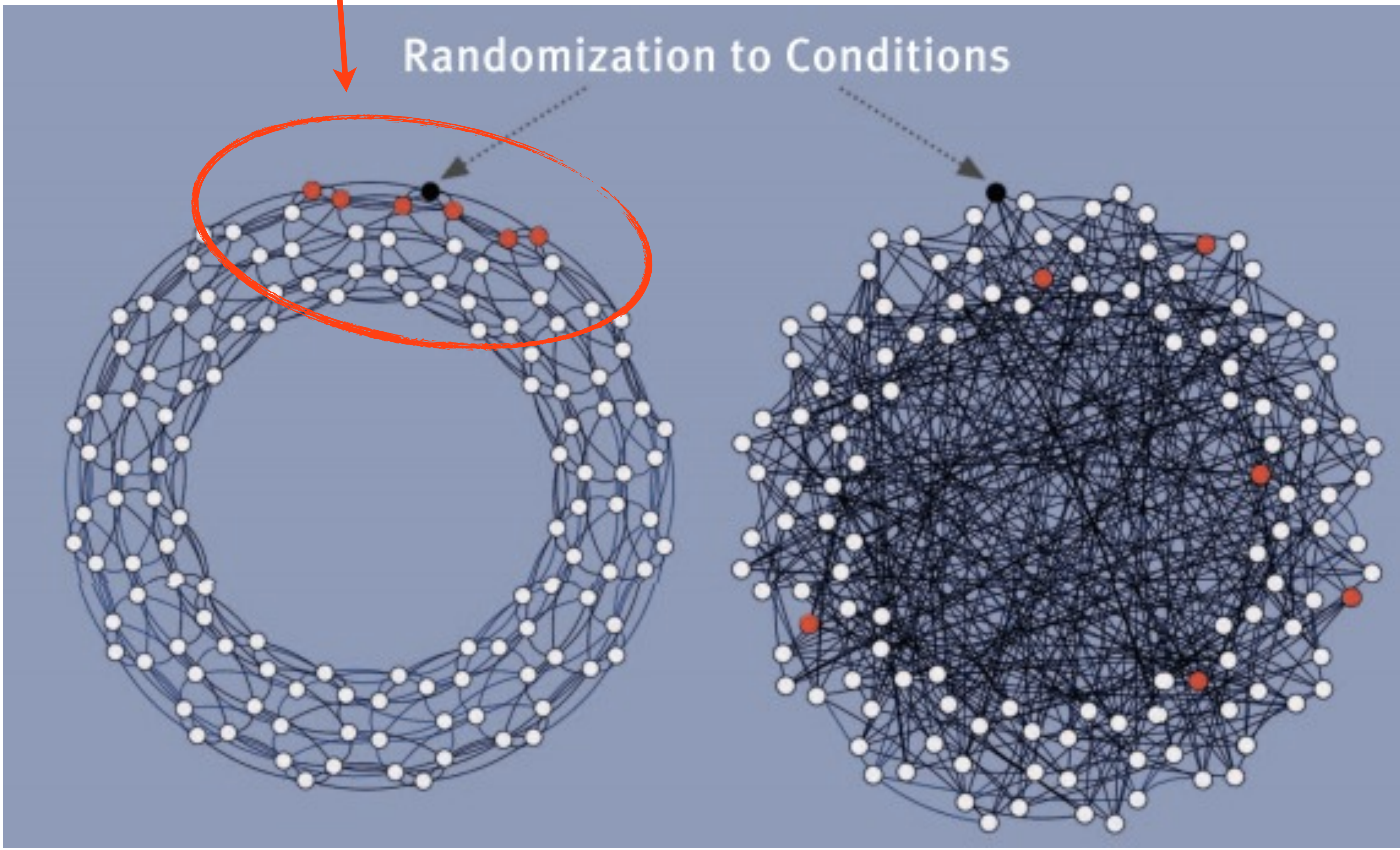

A
B
C
D
Fraction Adopted
Time (Days)


Randomization to Conditions


2.25
2.00
1.75
1.50
1.25
1.00
0.75
Hazard Ratio
2
3
4
Reinforcing Signals

**Multiple** exposures are crucial.

# Social reinforcement



# **Complex** Contagion

三人成虎

# 三人成虎

three

# 三人成虎

three

people

construct

三人成虎

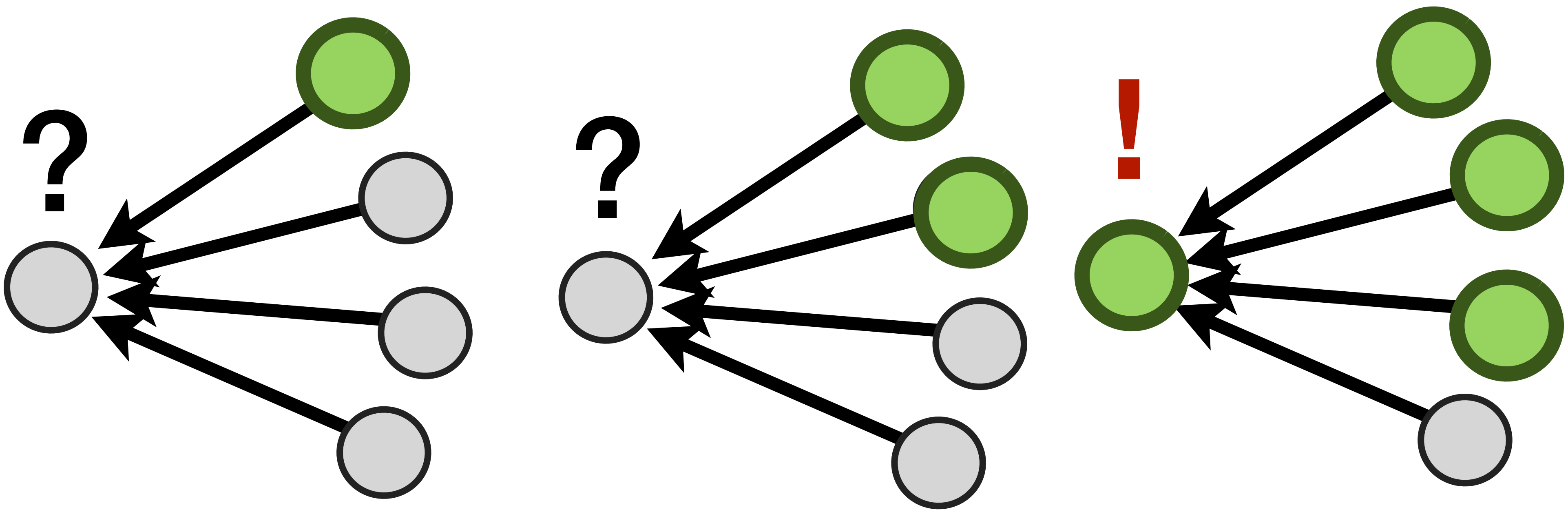
three

people

construct

三人成虎

three people tiger

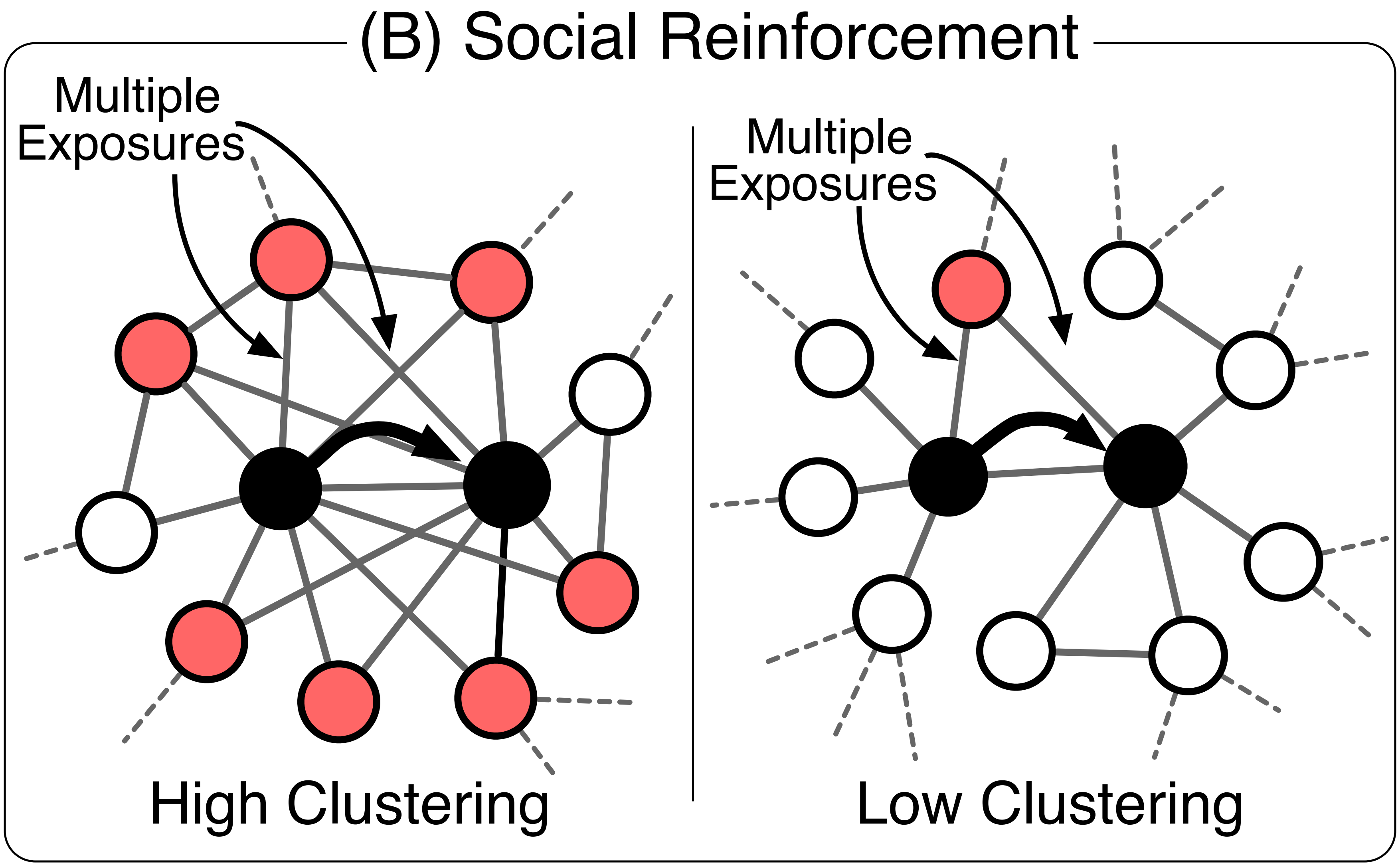


**Complex** Contagion needs multiple exposures.

&

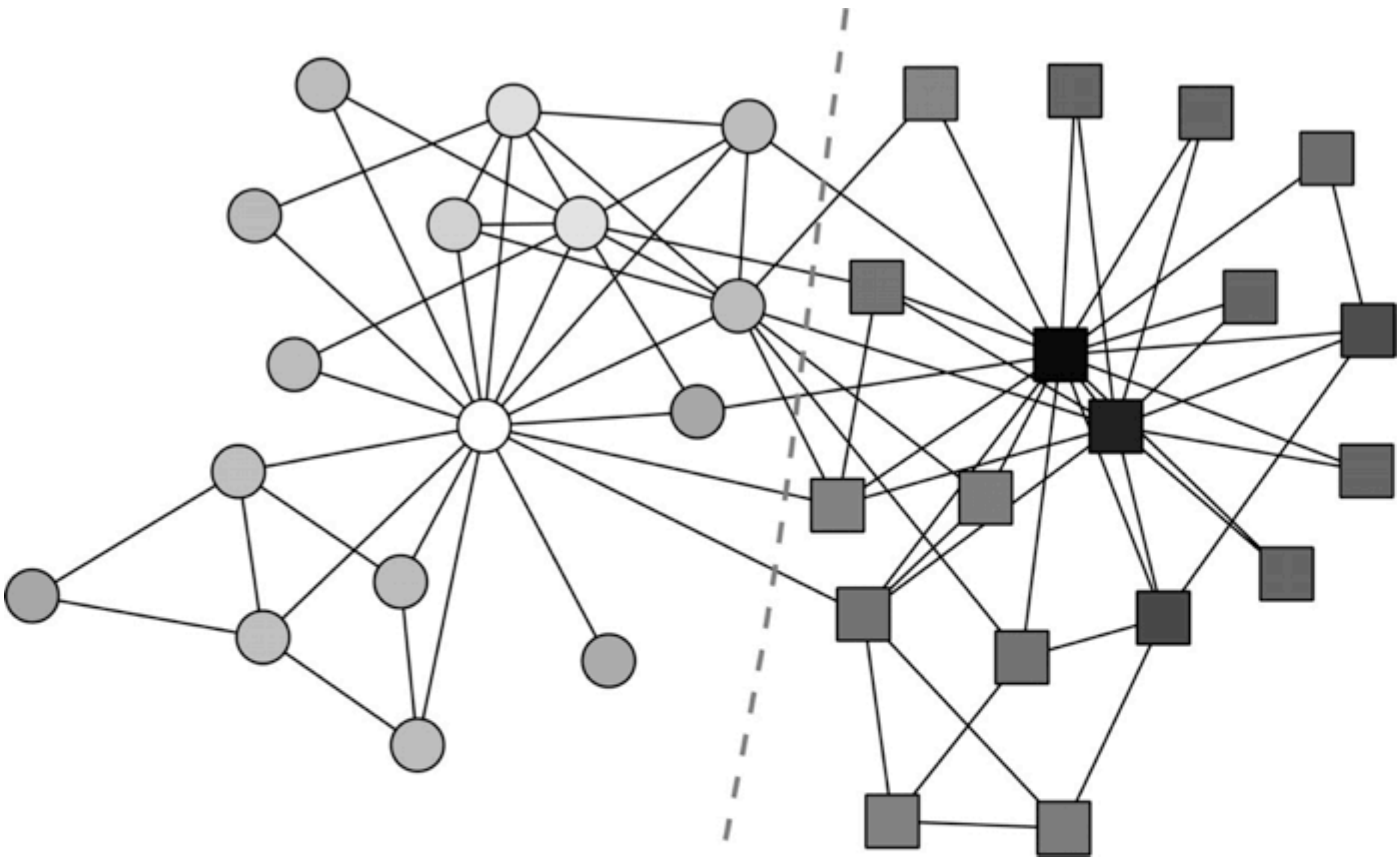
**Social** contagion seems to be **complex** contagions.

**Clustering** should be important.


(B) Social Reinforcement
Multiple Exposures
High Clustering
Multiple Exposures
Low Clustering

Highly clustered structure?

**Communities!**

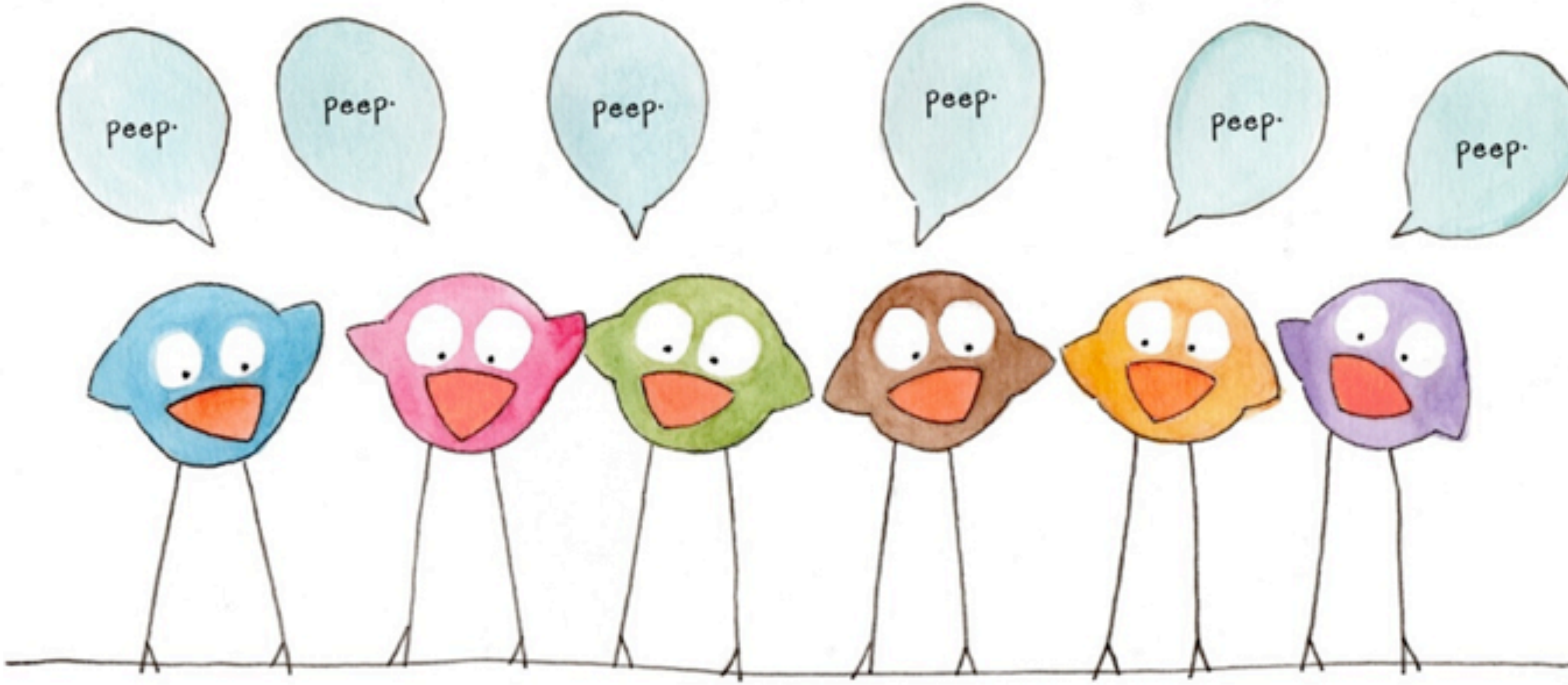


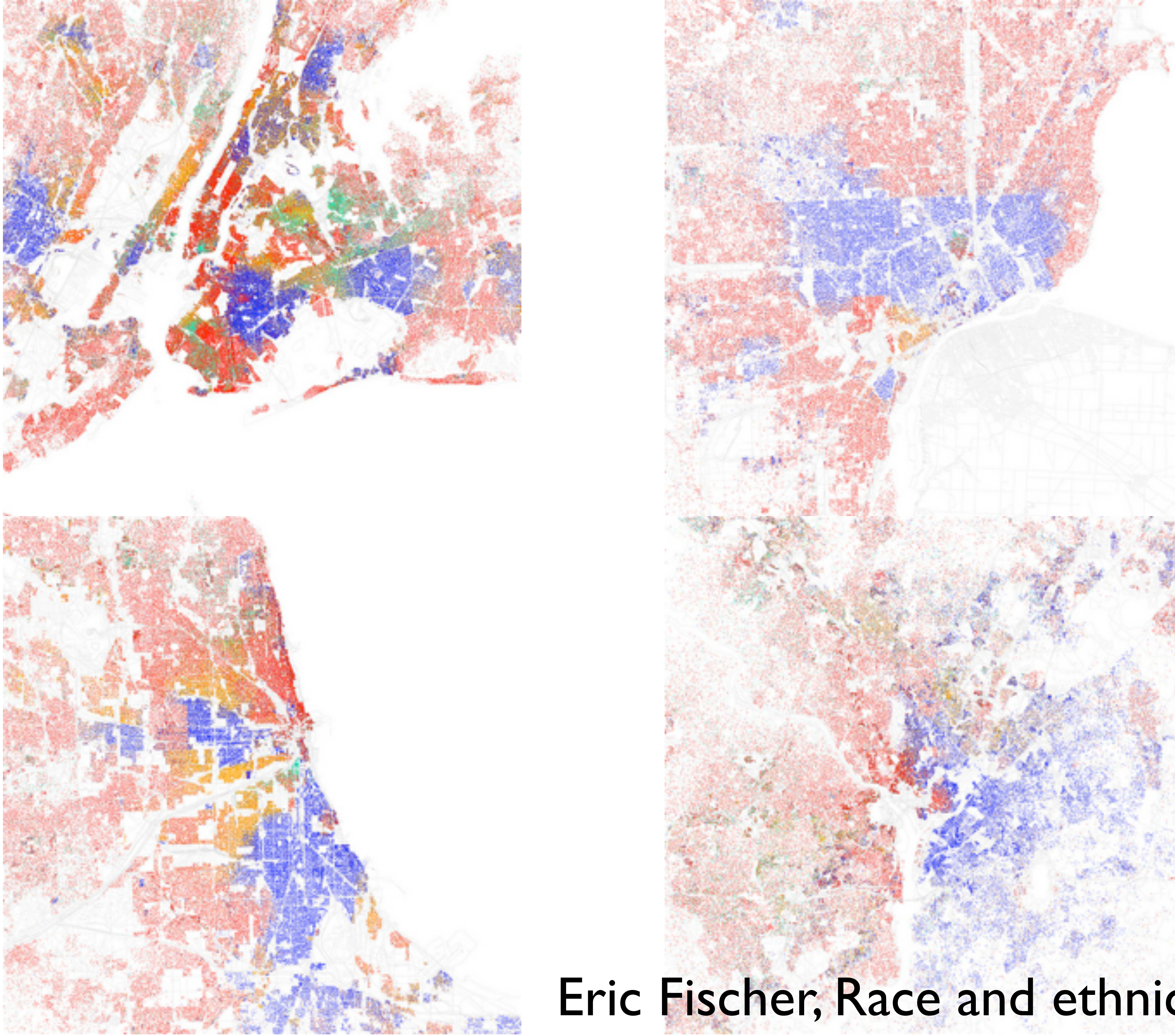
Newman, 2006

**Communities** should **enhance** the spread of **complex contagion.**

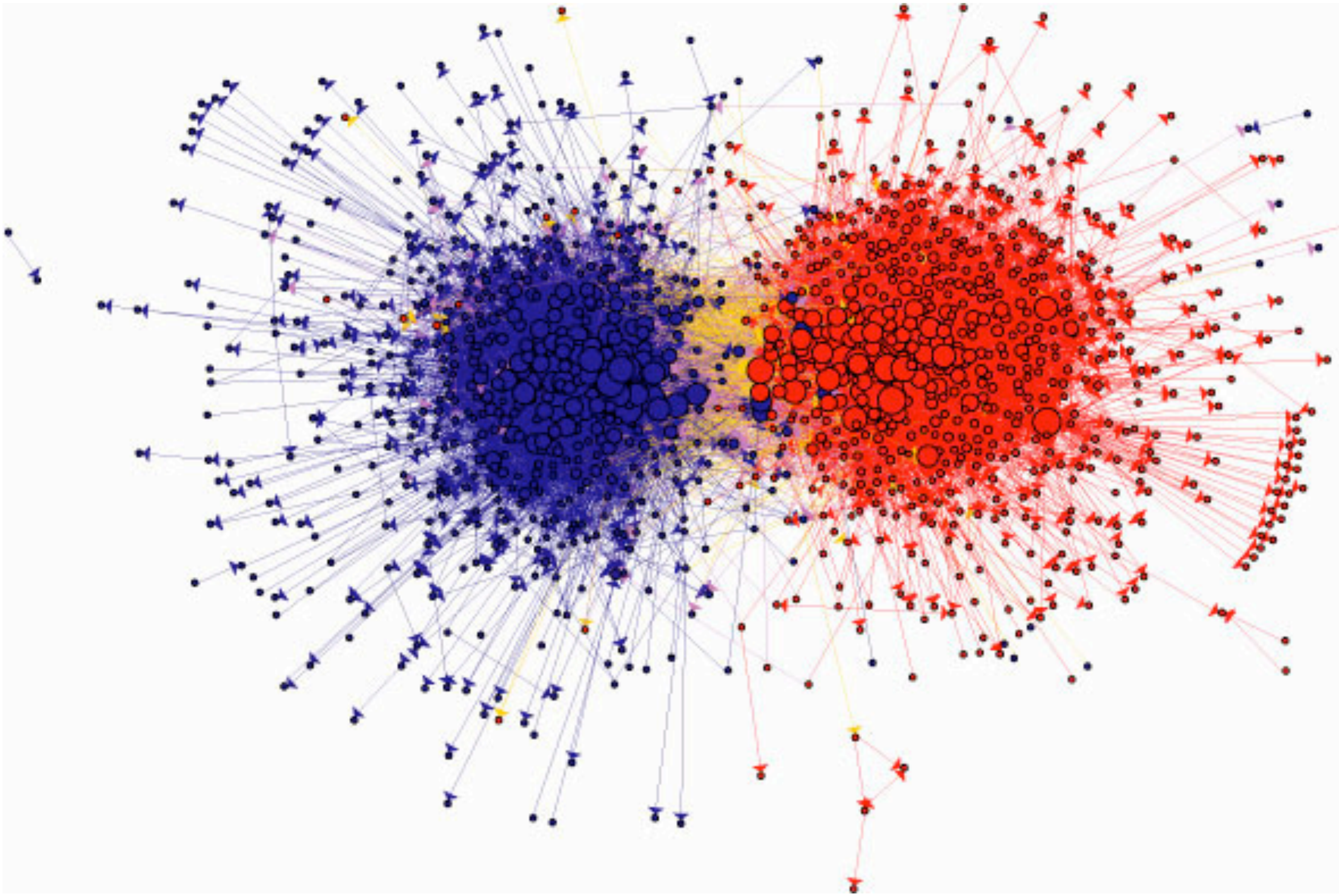
# Clue #2:
# **Homophily**

# “Birds of a feather flock together.”





Eric Fischer, Race and ethnicity



Adamic & Glance, 2005

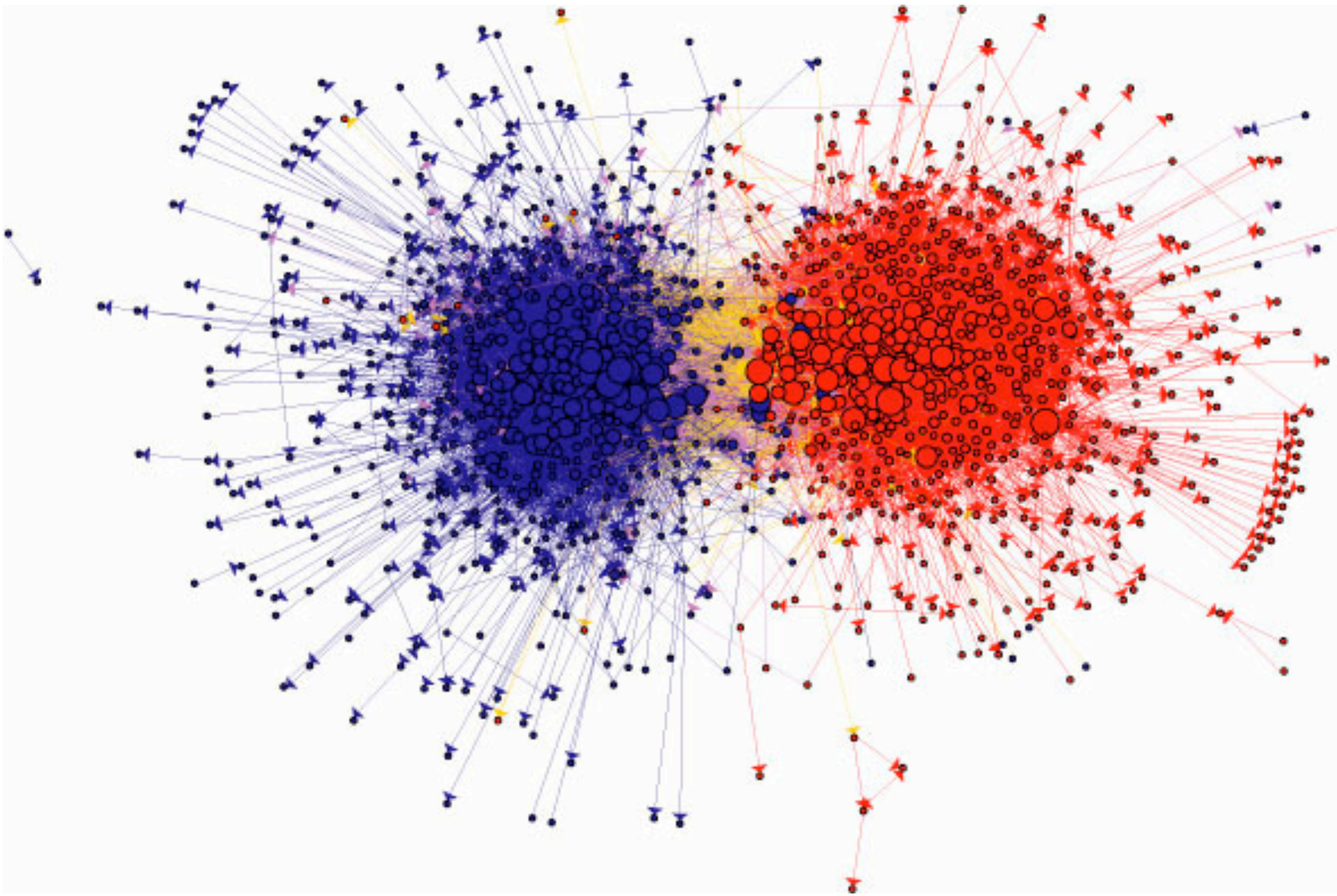
You are likely to share **similar interests** with your friends.

You are more likely to **adopt** something **from your social circles.**

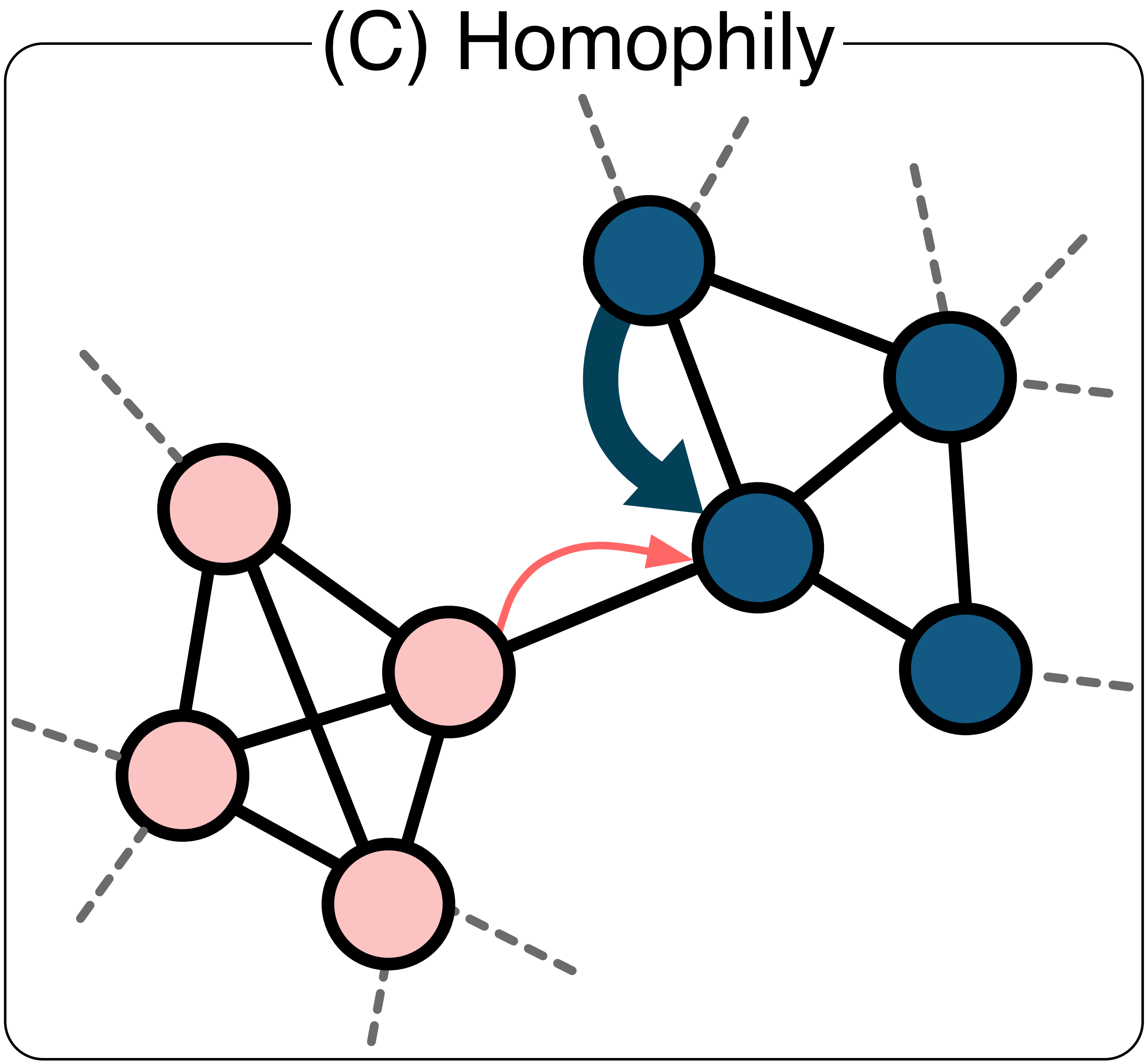
Again, **communities** become important.

**A community**

~ a common characteristic or shared interests

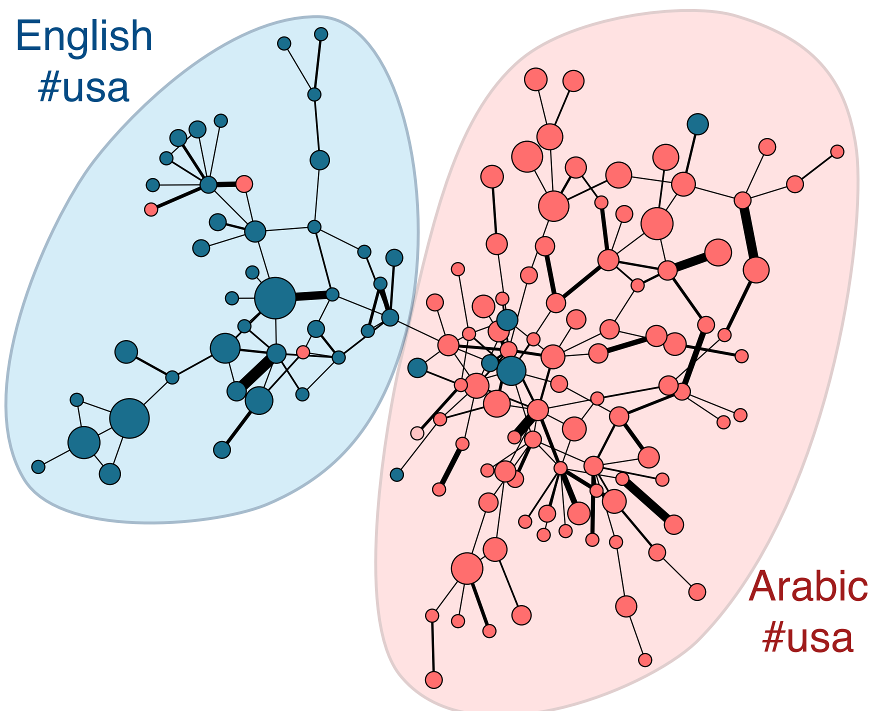


Adamic & Glance, 2005

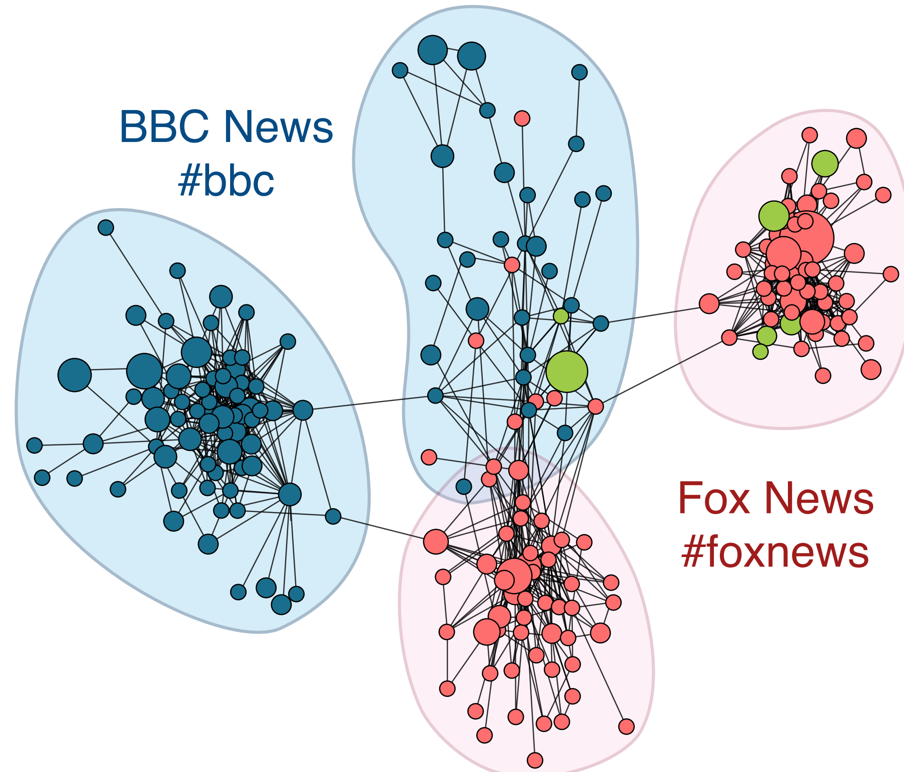

(C) Homophily

Trivial example:

Language and Country




Retweet Network

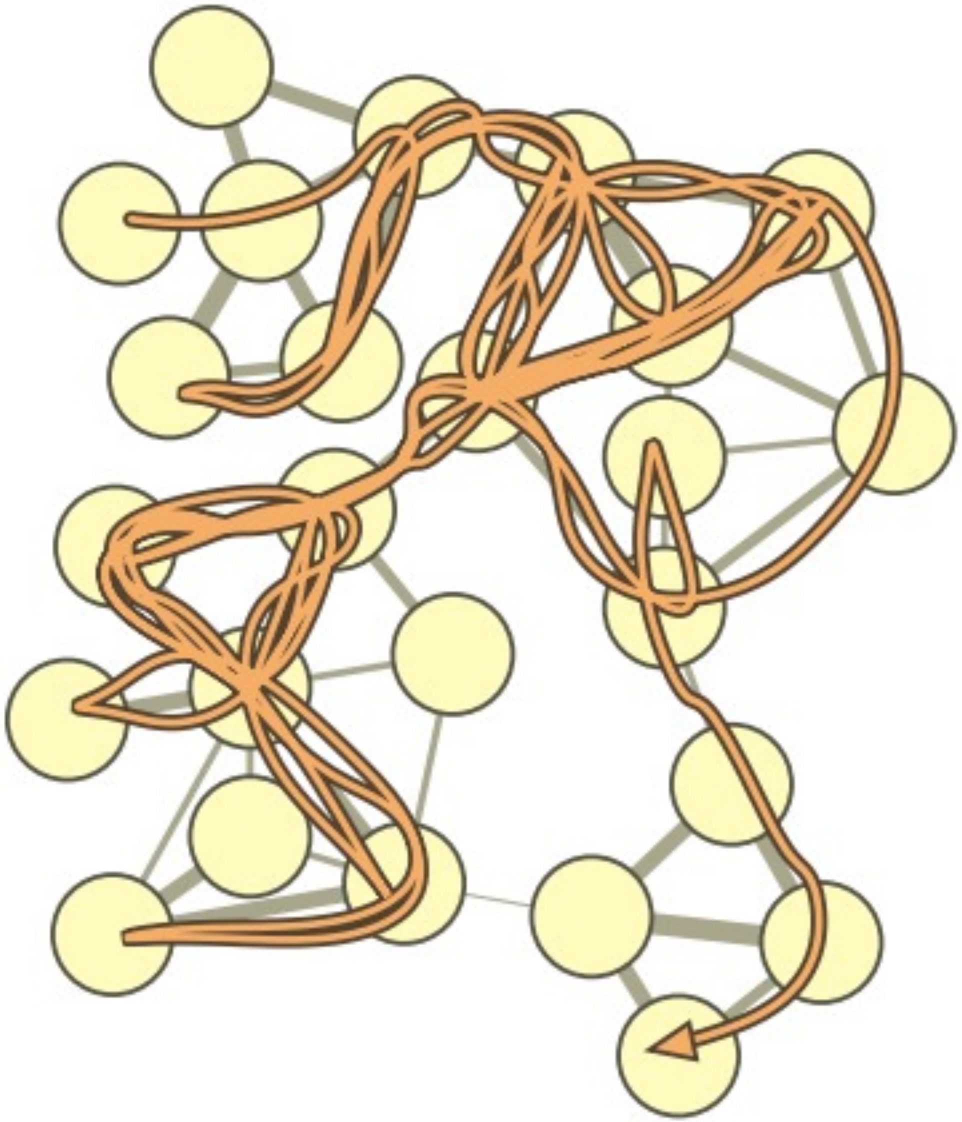



Follower Network

Both clues indicate that **Communities** should play a crucial role in **complex contagion**.

Communities **weakly trap simple** contagions.

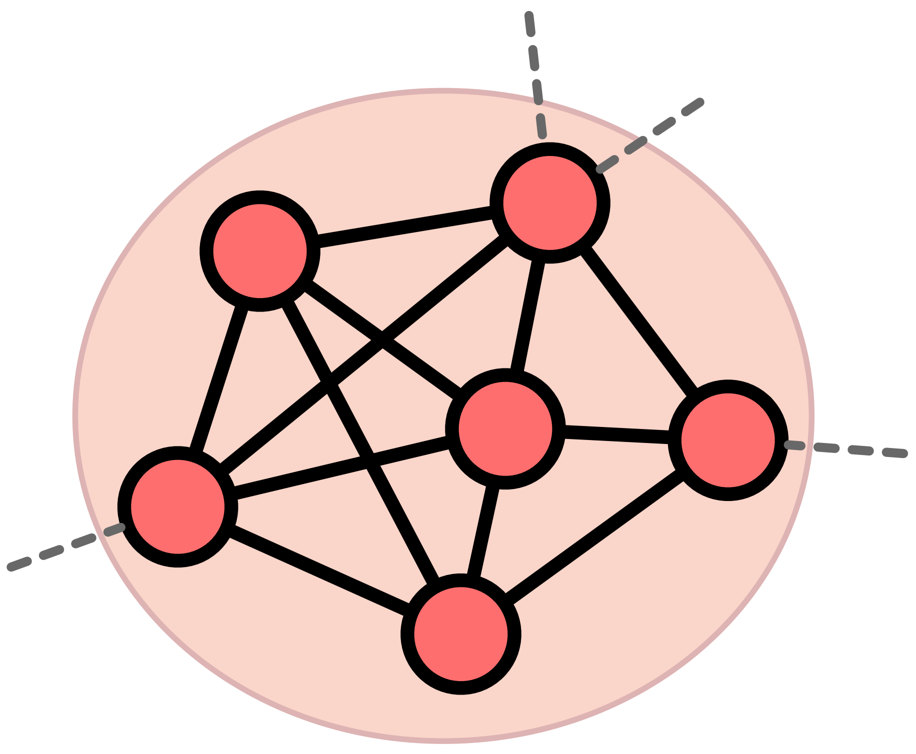
Communities **strongly trap complex** contagions.



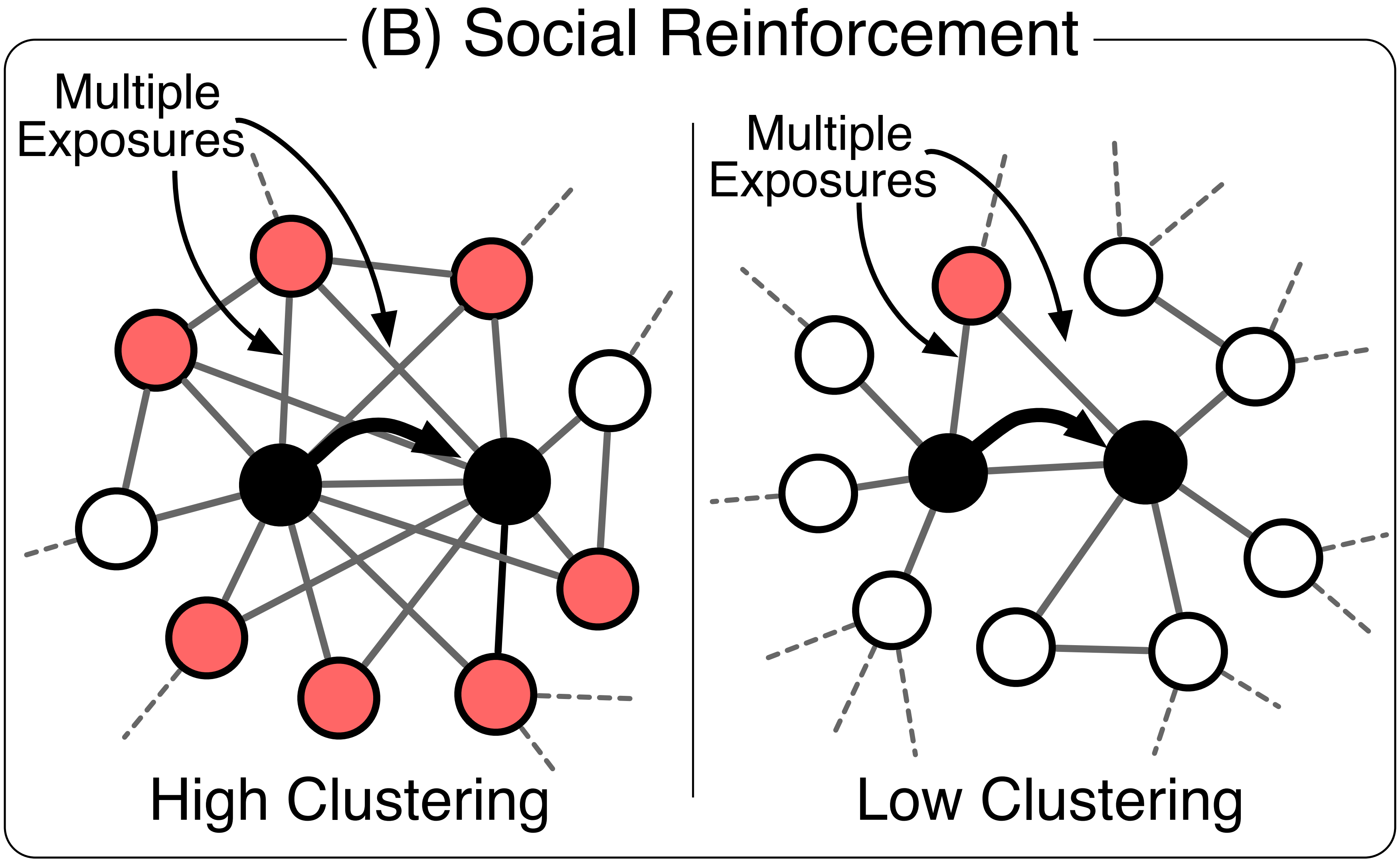
Communities:
traps for random walkers

Rosvall, Bergstrom, Lambiotte, ...

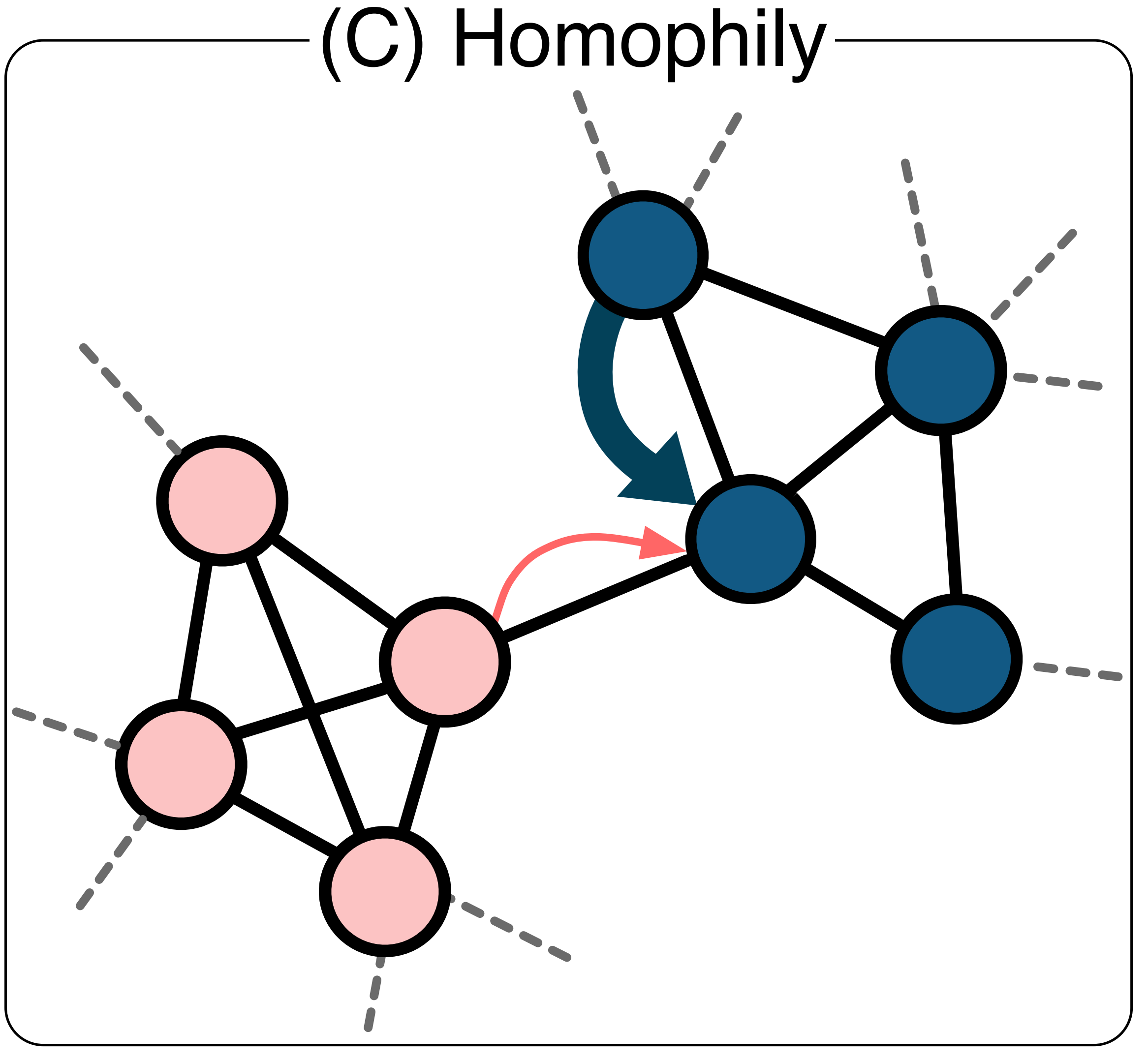
(A) Structural Trapping



Affects **both** simple and complex contagions.

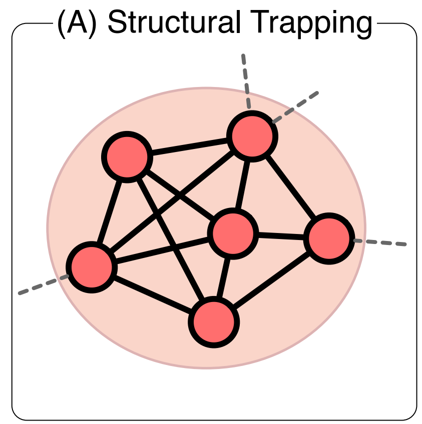



Affects complex contagions.

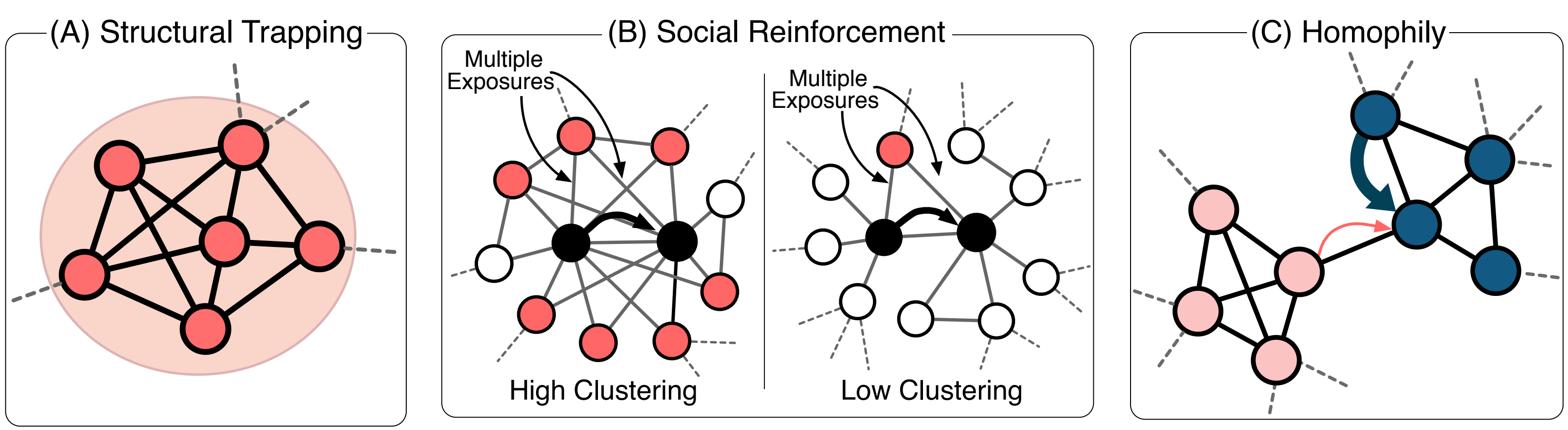



Affects complex contagions.

# Simple contagion

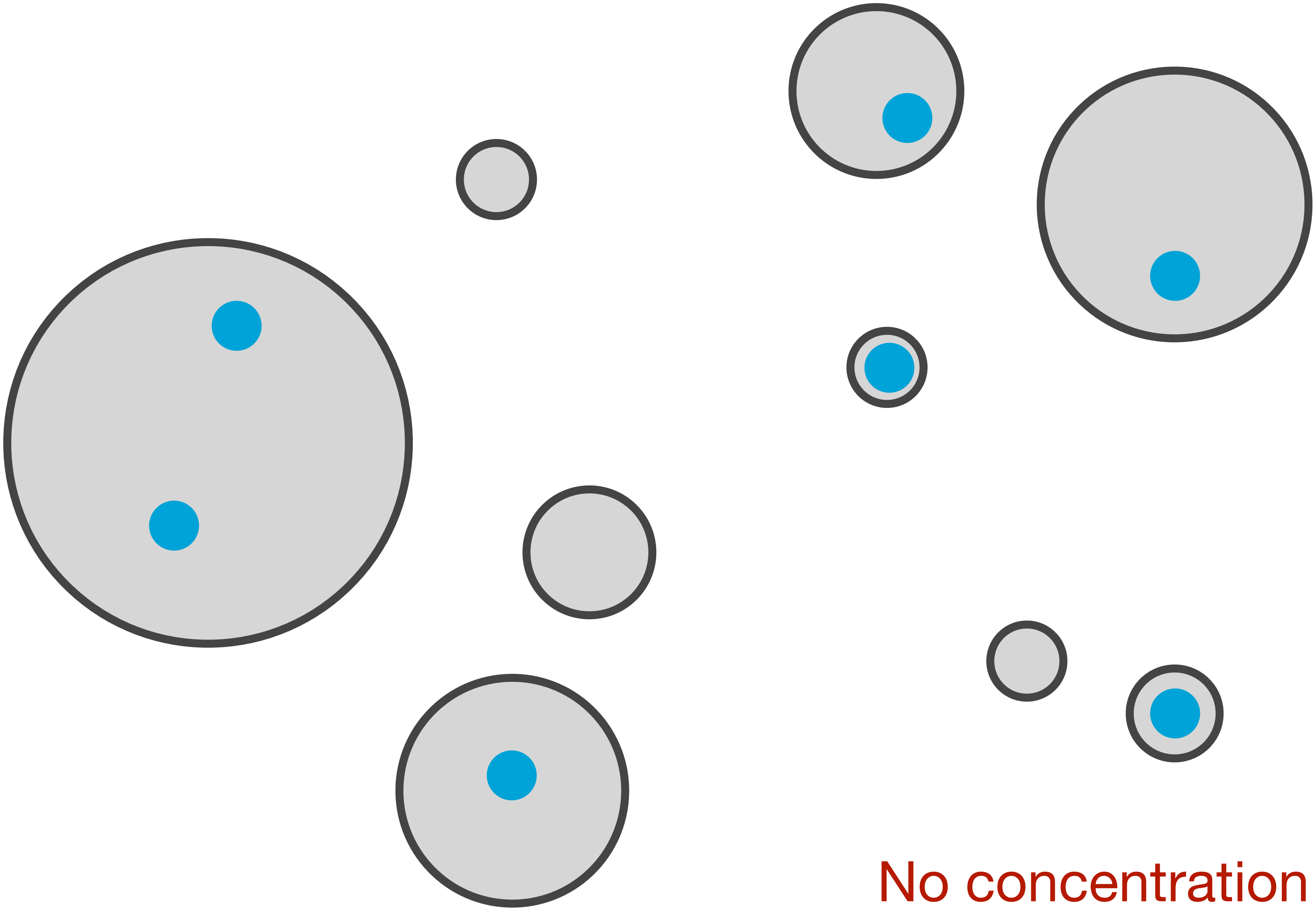



# Complex contagion

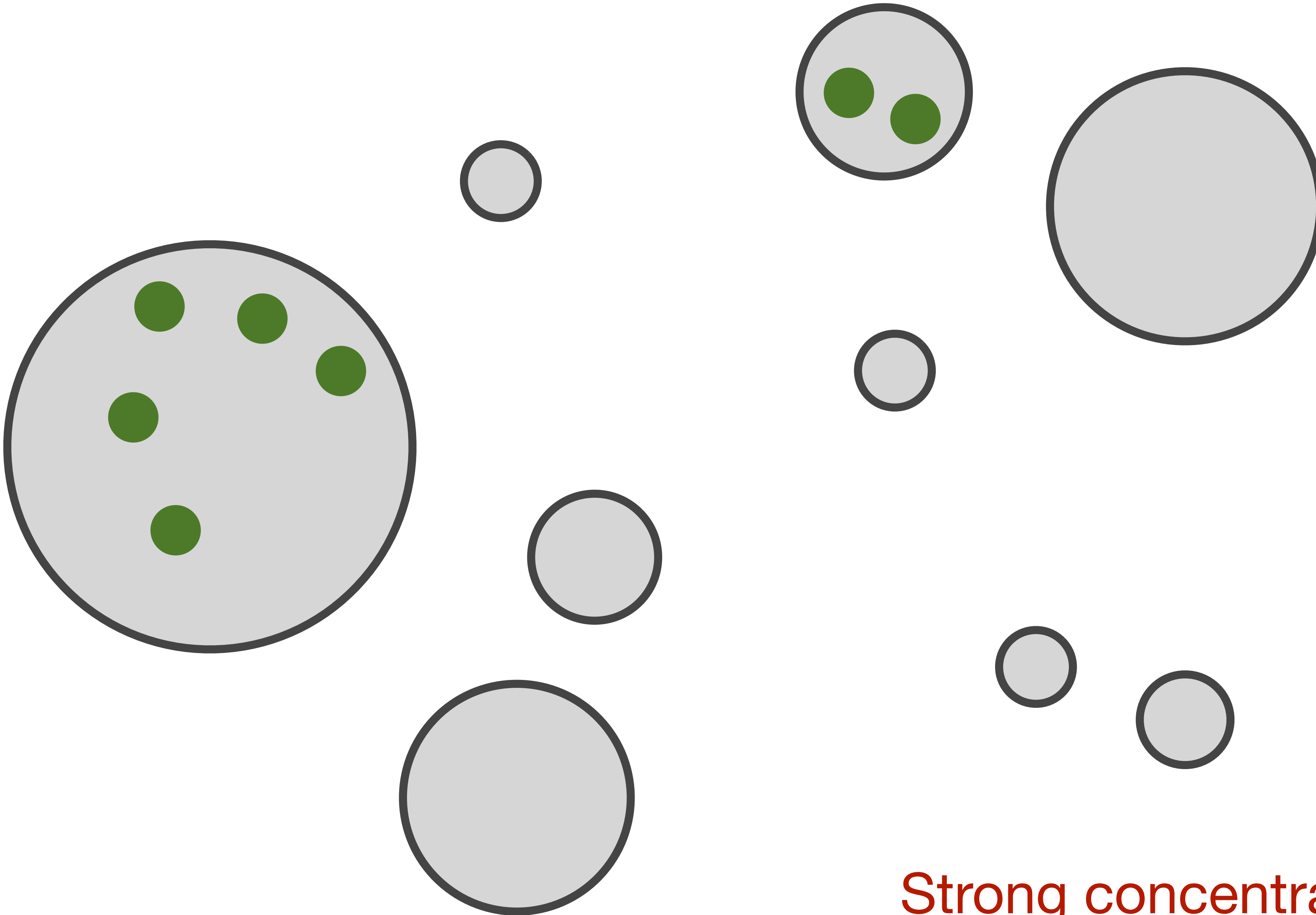


Communities **weakly trap simple** contagions.

Communities **strongly trap complex** contagions.



No concentration



Strong concentration

If our idea is correct,
then we will see
**strong concentration.**

In **complex contagion**, The edges in the communities should **transmit more information**.

# What about simple contagion?

**Traversing probability of an edge** from many events of simple contagion

~ that from many random walks

$P$: transition matrix

$P$: transition matrix

$$P^T \pi = \pi$$

$P$ : transition matrix $\qquad P^T \pi = \pi$

$$\begin{aligned} w^{\mathrm{rw}}(u,v) &= \pi_u p_{u\to v} + \pi_v p_{v\to u} \\ &= \frac{k(u)}{\sum_m k(m)} \frac{1}{k(u)} + \frac{k(v)}{\sum_m k(m)} \frac{1}{k(v)} \\ &= \frac{2}{\sum_m k(m)} \sim \text{const.} \end{aligned}$$

For **simple contagion**, we expect to see **no difference** between edges **inside** communities vs. ones **between** communities.

In **complex contagion**, The edges in the communities should **transmit more information**.

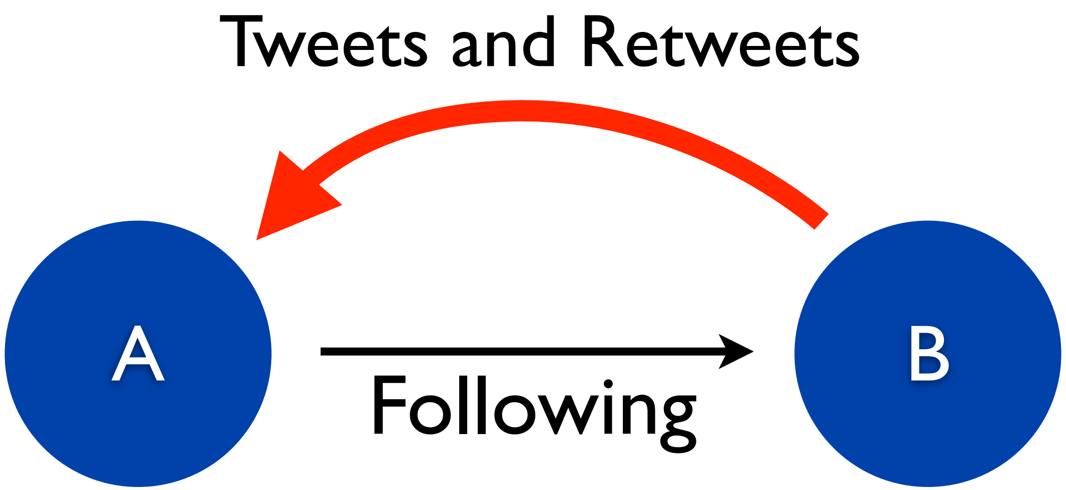
Then, why don't we measure the **concentration of memes** and **edge activities** regarding communities?

twitter

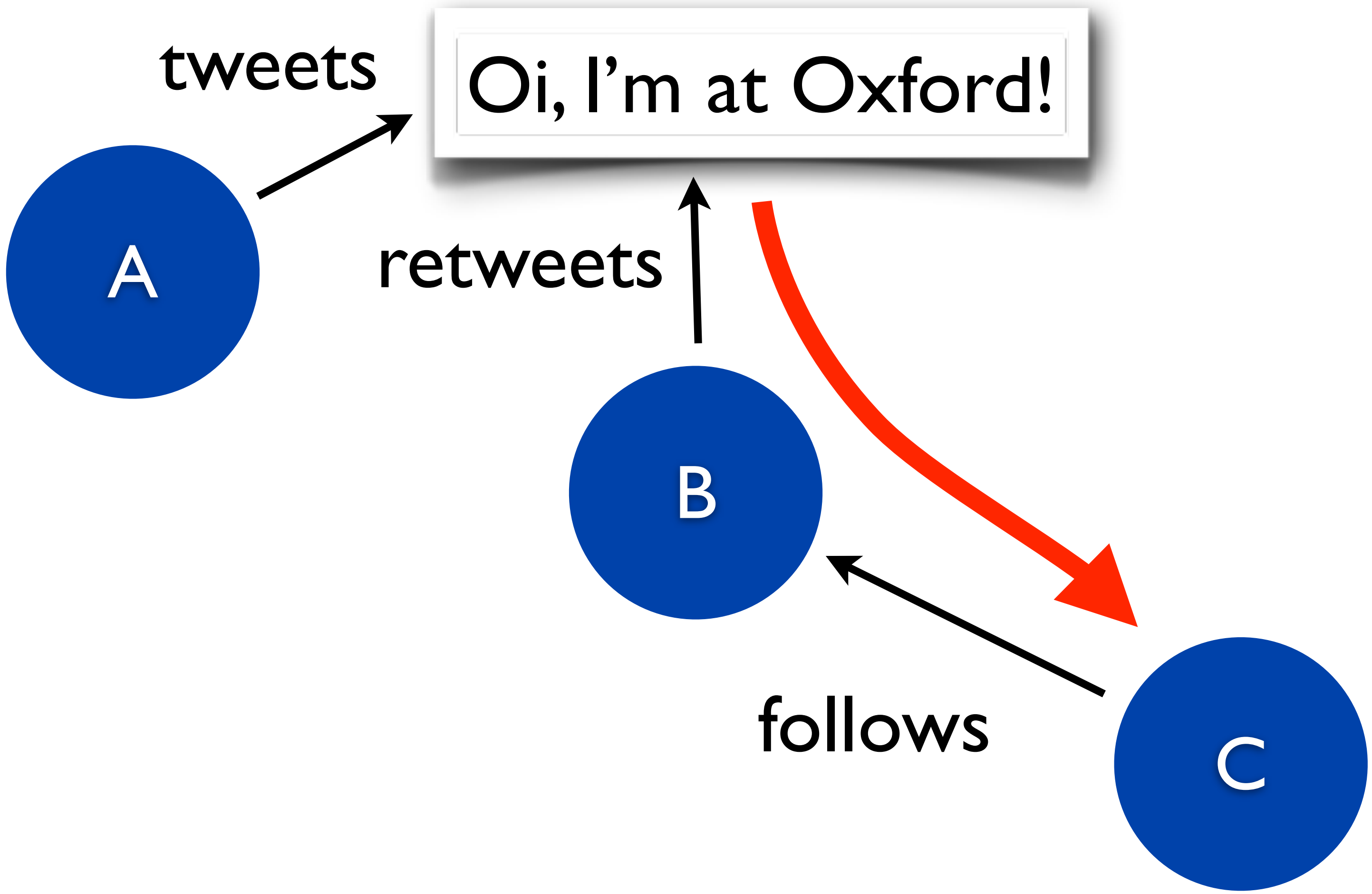
Provides data of both social networks and meme diffusion.

A multiplex, time-dependent network.

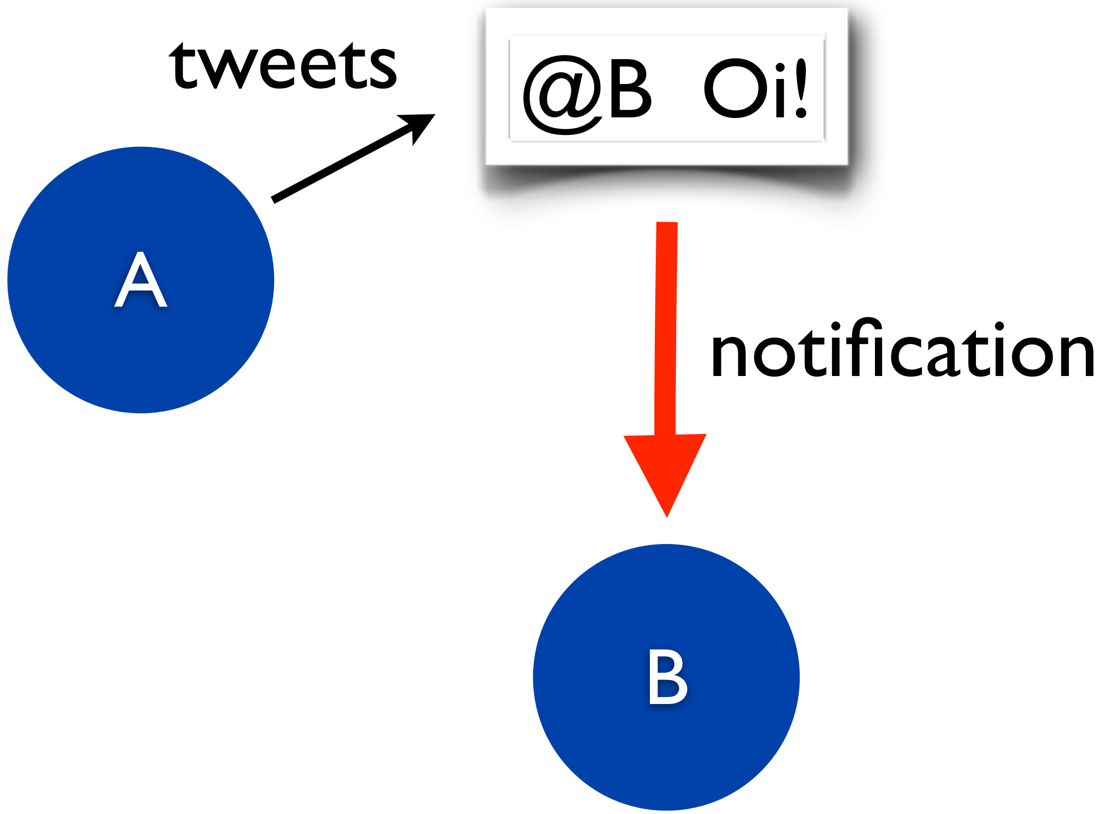
**Following, RT, mention**




If A **follows** B, B's tweets and retweets will appear in A's timeline




If B **retweets** a tweet, this tweets show up in the timelines of B's followers.



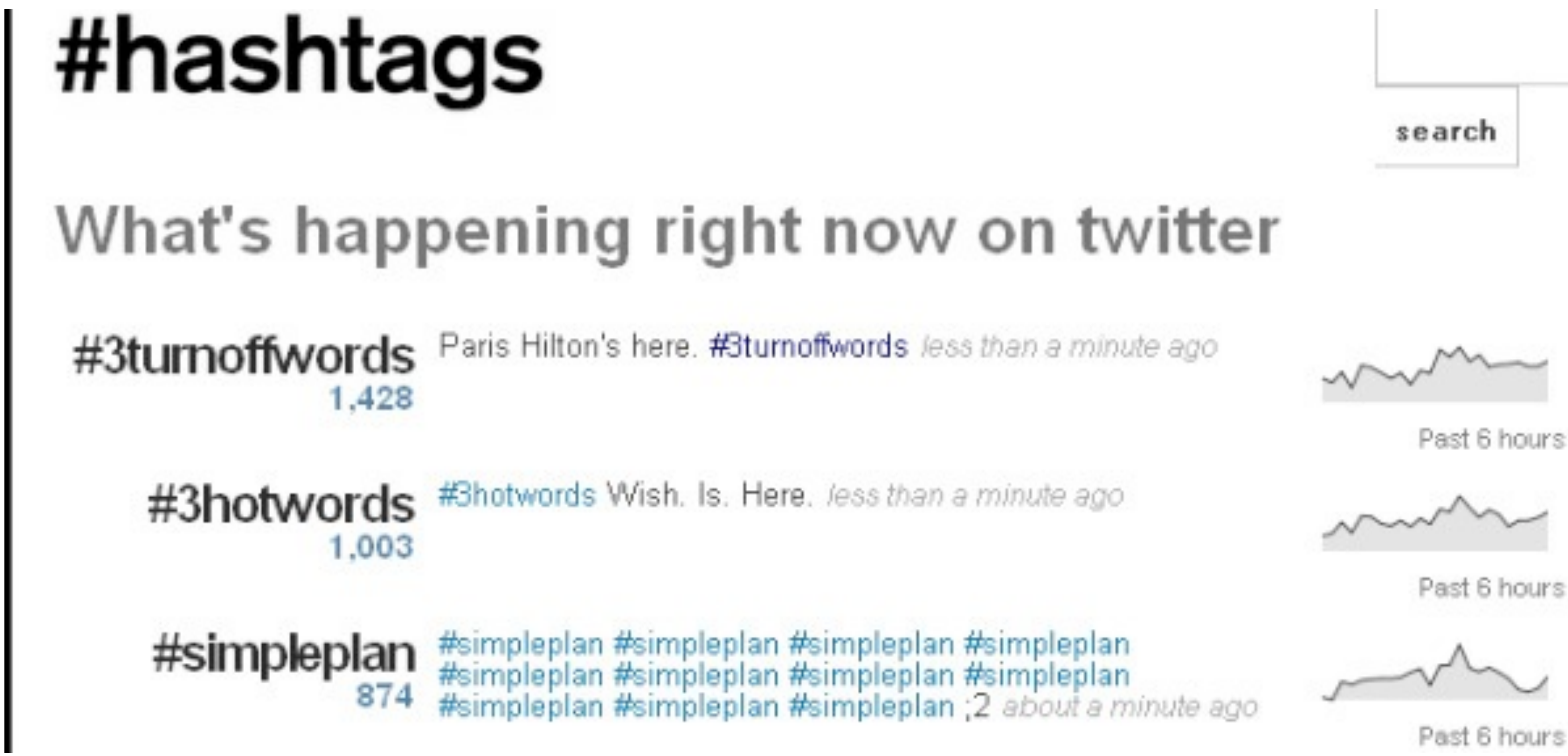

If A **mentions** B (with @B), B gets a notification about the tweet.

As an initial analysis, we constructed **three networks separately**.

**120 million tweets**
(Mar 24 – Apr 25, 2012)

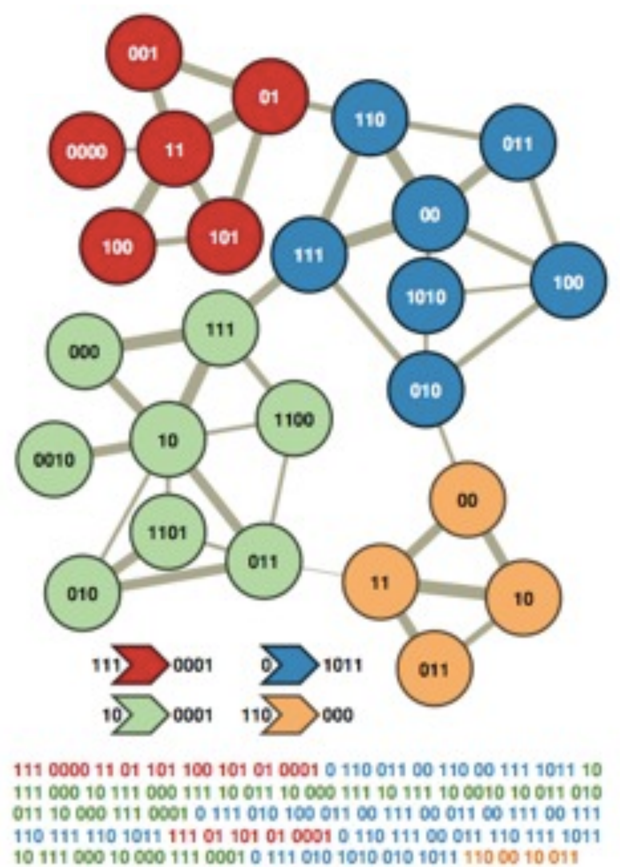
**600k** users, only reciprocal edges.

# Hashtag ~ Meme

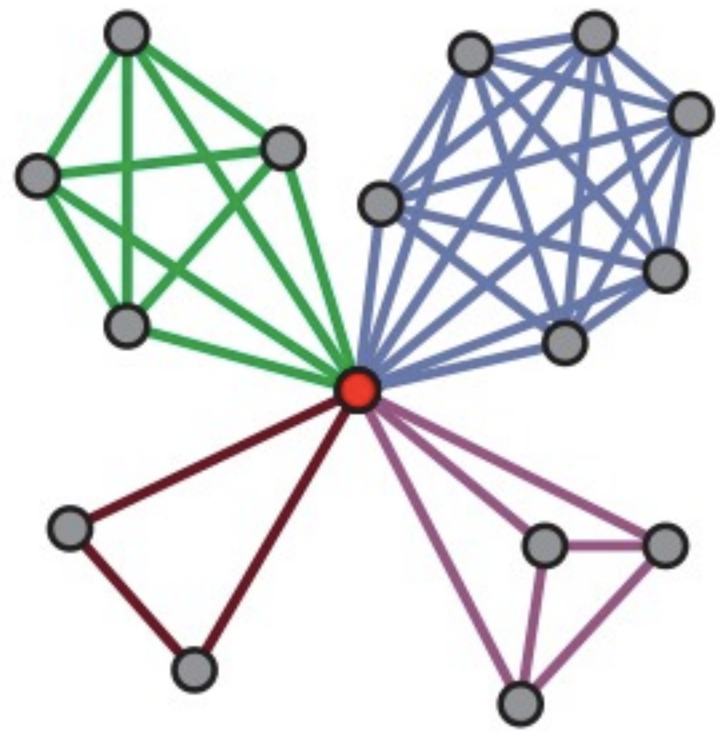


From 10 million hashtags, we pick only the ‘**new**’ hashtags (fewer than 20 tweets in the previous month).

# Two community detection methods



Infomap (Rosvall & Bergstrom, 2008)



Link clustering (Ahn, Bagrow, Lehmann, 2010)

Didn’t use edge weights when detecting communities.

| | | Network types | | |
|---|---|---|---|---|
| | | Retweet | Mention | Follower |
| | Number of nodes | 300,197 | 374,829 | 595,460 |
| | Number of edges | 598,487 | 1,048,818 | 14,273,311 |
| | Avg. clustering coefficient | 0.0902 | 0.1284 | 0.1972 |
| InfoMap | Number of communities | 14,144 | 14,222 | 6,360 |
| | Node coverage | 99.86% | 99.72% | 99.72% |
| LinkComm | Number of communities | 57,317 | 97,198 | 321,774 |
| | Node coverage | 48.42% | 67.23% | 47.62% |

Are memes **complex contagions**?

If it's **complex contagion**,
The edges inside
communities should
**transmit more**
**information.**

Two types of edges in the **following graph**.

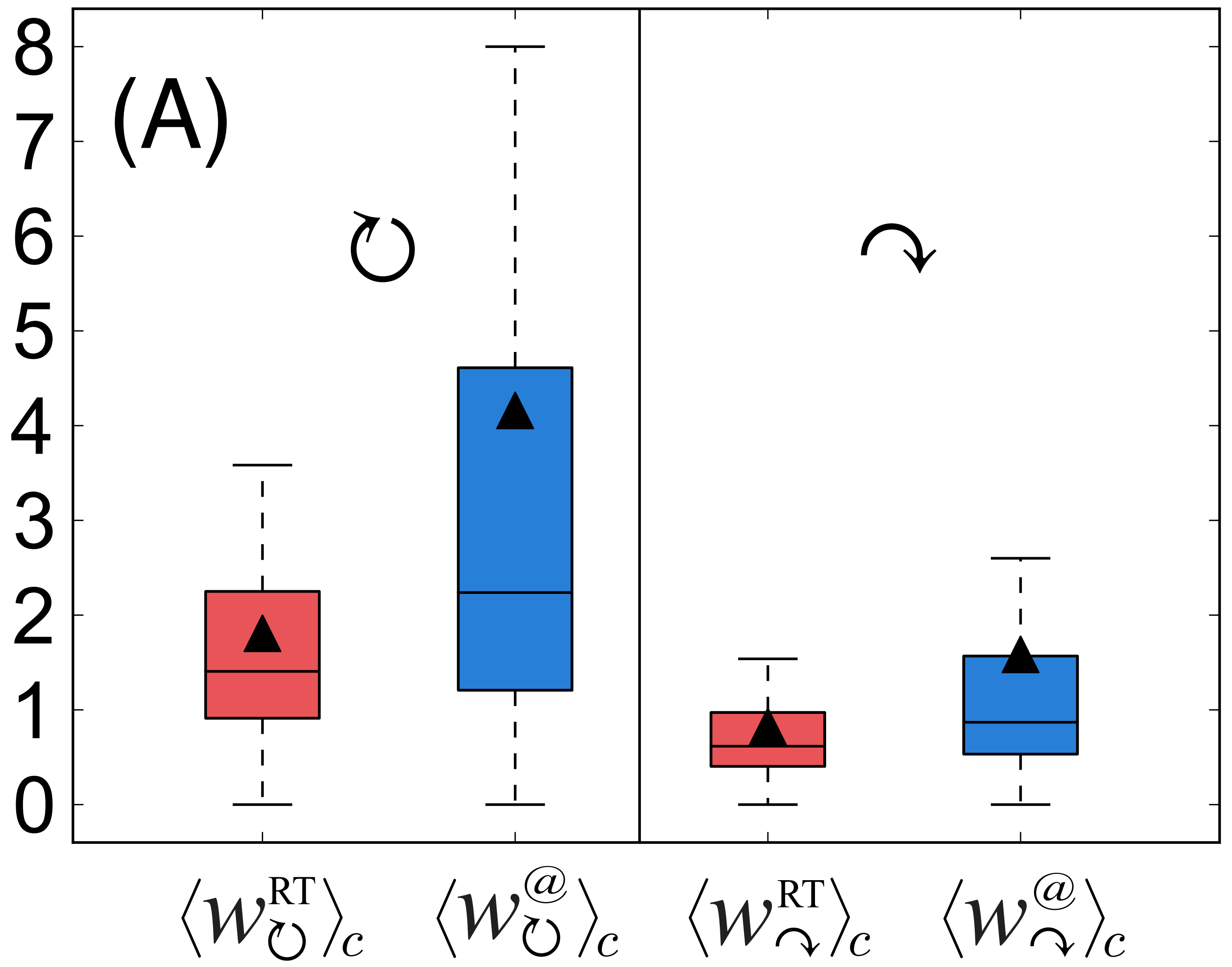
Intra-edges: $E_{\circlearrowright}$

Inter-edges: $E_{\curvearrowright}$

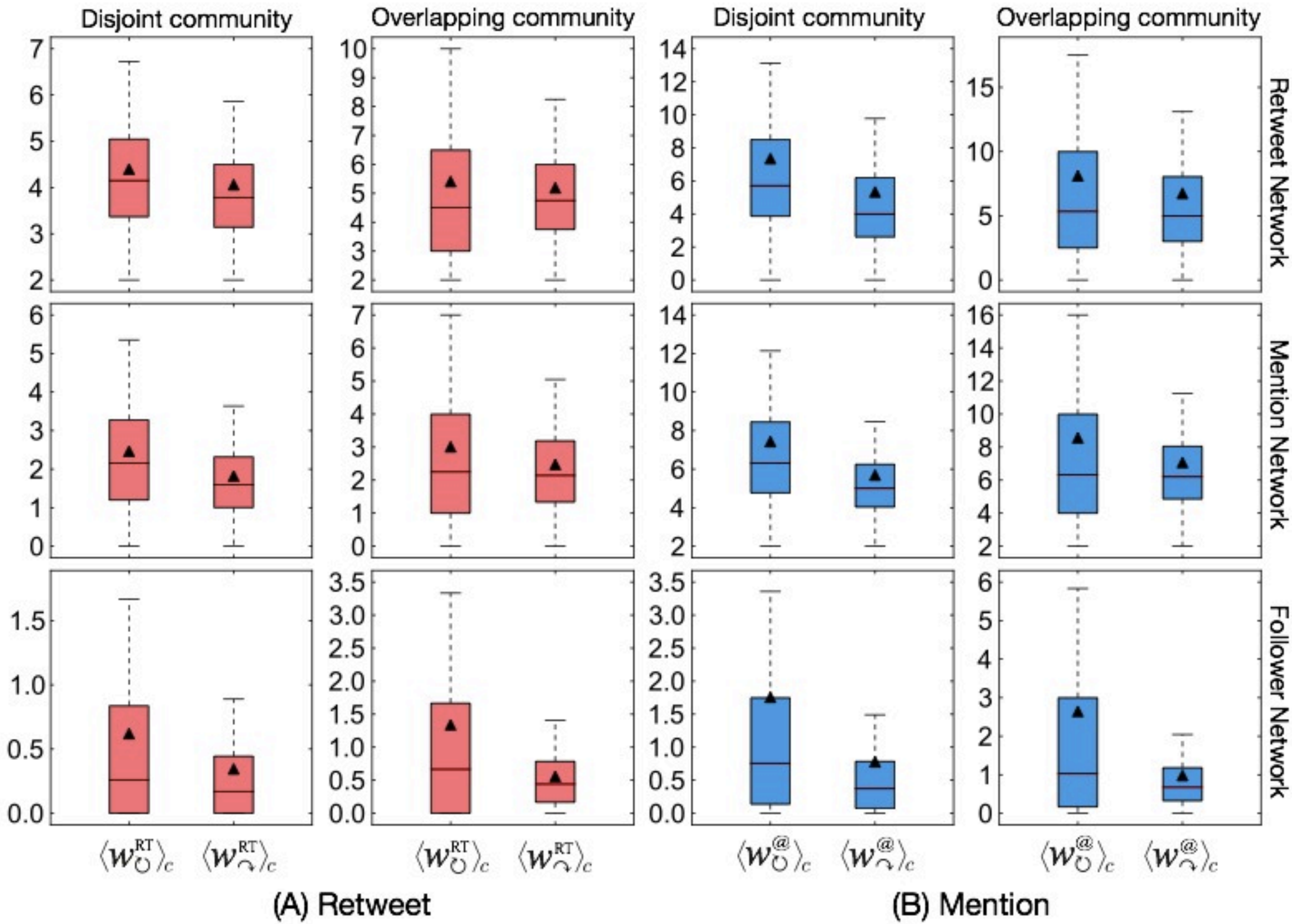
$$\langle w_{\circlearrowright} \rangle_c = \frac{1}{|E^c_{\circlearrowright}|} \sum_{(u,v)\in E^c_{\circlearrowright}} w(u,v)$$

$$\langle w_{\curvearrowright} \rangle_c = \frac{1}{|E^c_{\curvearrowright}|} \sum_{(u,v)\in E^c_{\curvearrowright}} w(u,v)$$

For each type of edges, we measure the **# of retweets and mentions through** the edges.



(A)
0
1
2
3
4
5
6
7
8
$\langle W^{\mathrm{RT}}_{\circlearrowright} \rangle_c$
$\langle W^{@}_{\circlearrowright} \rangle_c$
$\langle W^{\mathrm{RT}}_{\curvearrowright} \rangle_c$
$\langle W^{@}_{\curvearrowright} \rangle_c$



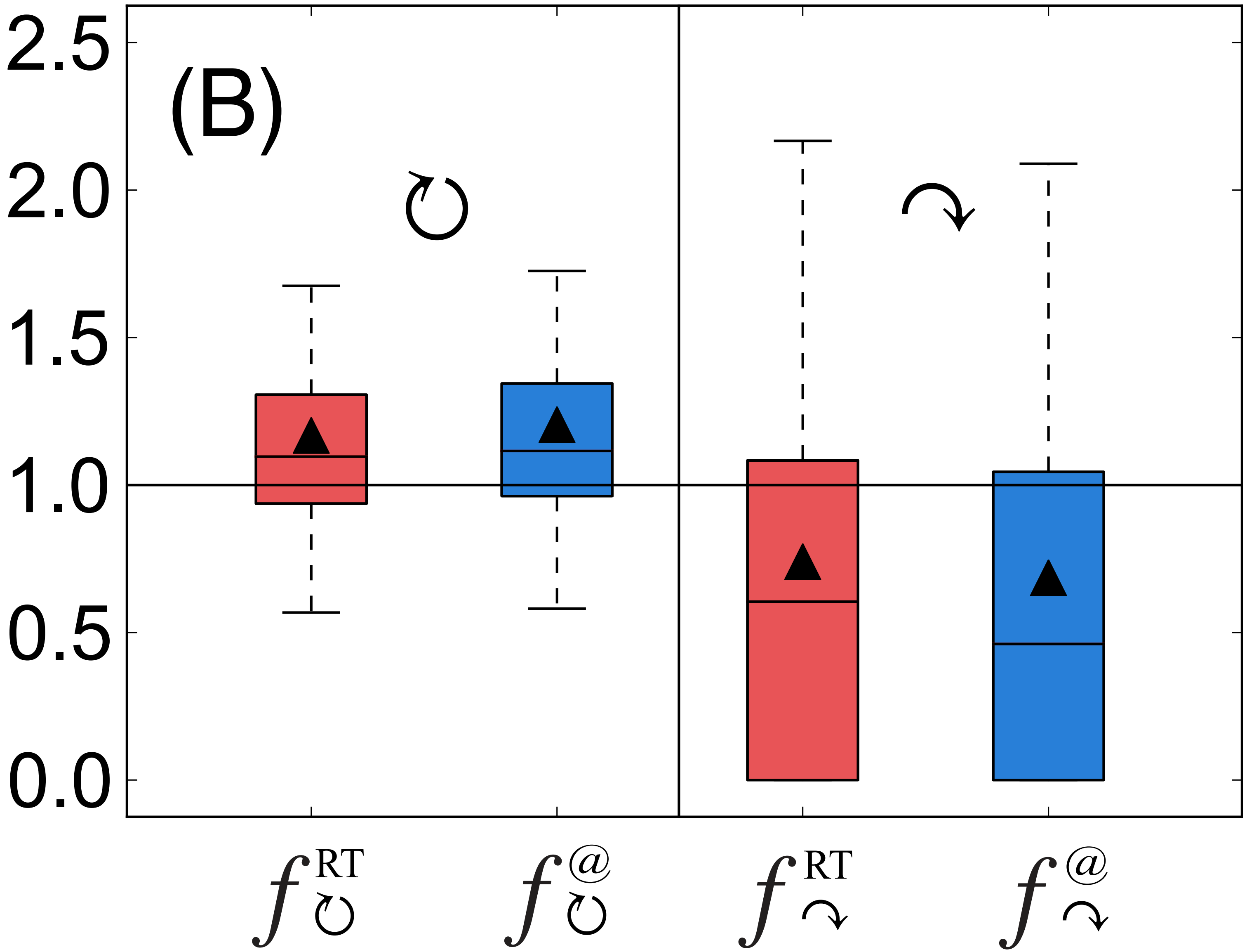
Disjoint community
Overlapping community
Disjoint community
Overlapping community
Retweet Network
Mention Network
Follower Network
$\langle w^{\mathrm{RT}}_{\circlearrowleft} \rangle_c$
$\langle w^{\mathrm{RT}}_{\curvearrowright} \rangle_c$
$\langle w^{@}_{\circlearrowleft} \rangle_c$
$\langle w^{@}_{\curvearrowright} \rangle_c$
(A) Retweet
(B) Mention

# Fraction of interactions for each person

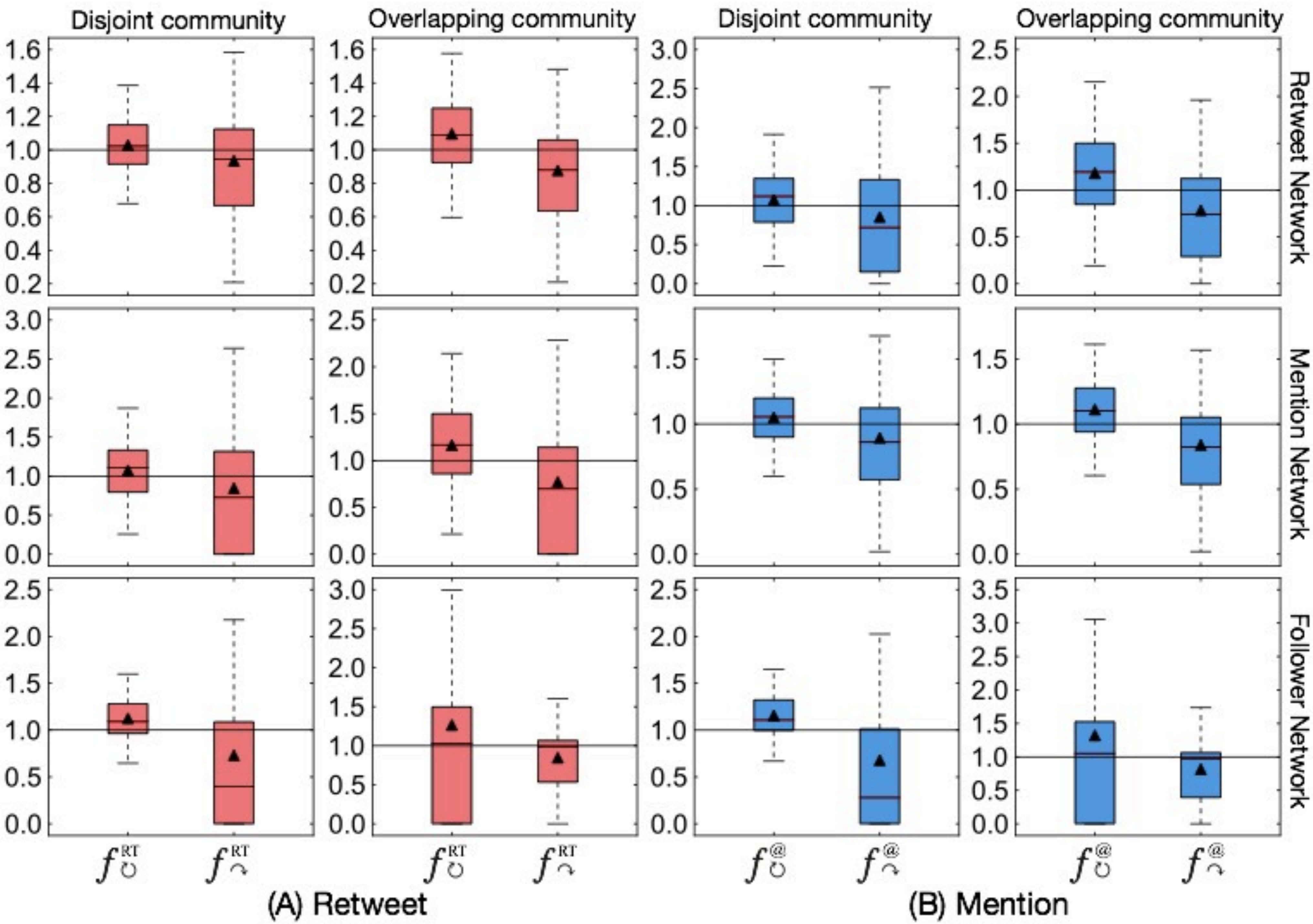
$$f_{\circlearrowright}(u) = \frac{\frac{1}{k_{\circlearrowright}(u)} \sum_{(u,v)\in E_{\circlearrowright}} w(u,v)}{\frac{1}{k(u)} \sum_{(u,v)\in E} w(u,v)}$$

$$f_{\curvearrowright}(u) = \frac{\frac{1}{k_{\curvearrowright}(u)} \sum_{(u,v)\in E_{\curvearrowright}} w(u,v)}{\frac{1}{k(u)} \sum_{(u,v)\in E} w(u,v)}$$

$$k_{\circlearrowright}(u) = |\{v \mid (u,v) \in E_{\circlearrowright}\}|$$
$$k_{\curvearrowright}(u) = |\{v \mid (u,v) \in E_{\curvearrowright}\}|$$
$$k(u) = k_{\circlearrowright}(u) + k_{\curvearrowright}(u).$$



(B)
2.5
2.0
1.5
1.0
0.5
0.0
$f^{RT}_{\circlearrowright}$
$f^{@}_{\circlearrowright}$
$f^{RT}_{\curvearrowright}$
$f^{@}_{\curvearrowright}$



Disjoint community
Overlapping community
Disjoint community
Overlapping community
Retweet Network
Mention Network
Follower Network
$f^{RT}_{\circlearrowright}$
$f^{RT}_{\curvearrowright}$
$f^{@}_{\circlearrowright}$
$f^{@}_{\curvearrowright}$
(A) Retweet
(B) Mention

Indeed, we see more activities **within** communities**.**

How do we measure "**concentration**"?

# We need models to compare

| | Network effect | Social reinforcement | Homophily | |
|---|---|---|---|---|
| M1 | | | | Random sampling |
| M2 | O | | | Simple cascade |
| M3 | O | O | | Social reinforcement |
| M4 | O | | O | Homophily |

$r(h)$ the proportion of **tweets** produced in the dominant community

$r(h)$ the proportion of **tweets** produced in the dominant community

$g(h)$ the proportion of **users** adopted in the dominant community

$r(h)$ the proportion of **tweets** produced in the dominant community

$g(h)$ the proportion of **users** adopted in the dominant community

$H^t(h)$ **Tweet** entropy in terms of communities

$r(h)$ the proportion of **tweets** produced in the dominant community

$g(h)$ the proportion of **users** adopted in the dominant community

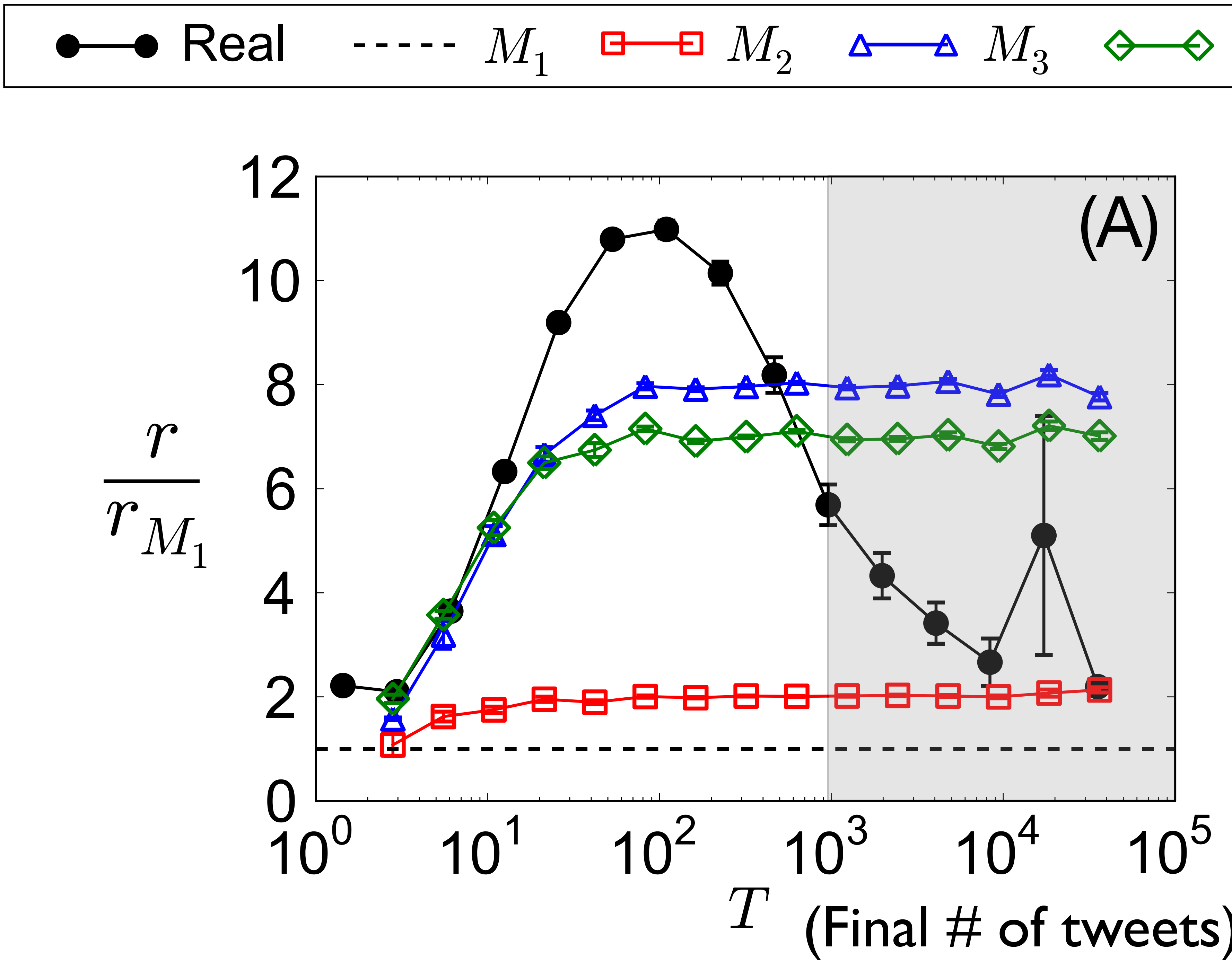
$H^t(h)$ **Tweet** entropy in terms of communities

$H^u(h)$ **User** entropy in terms of communities

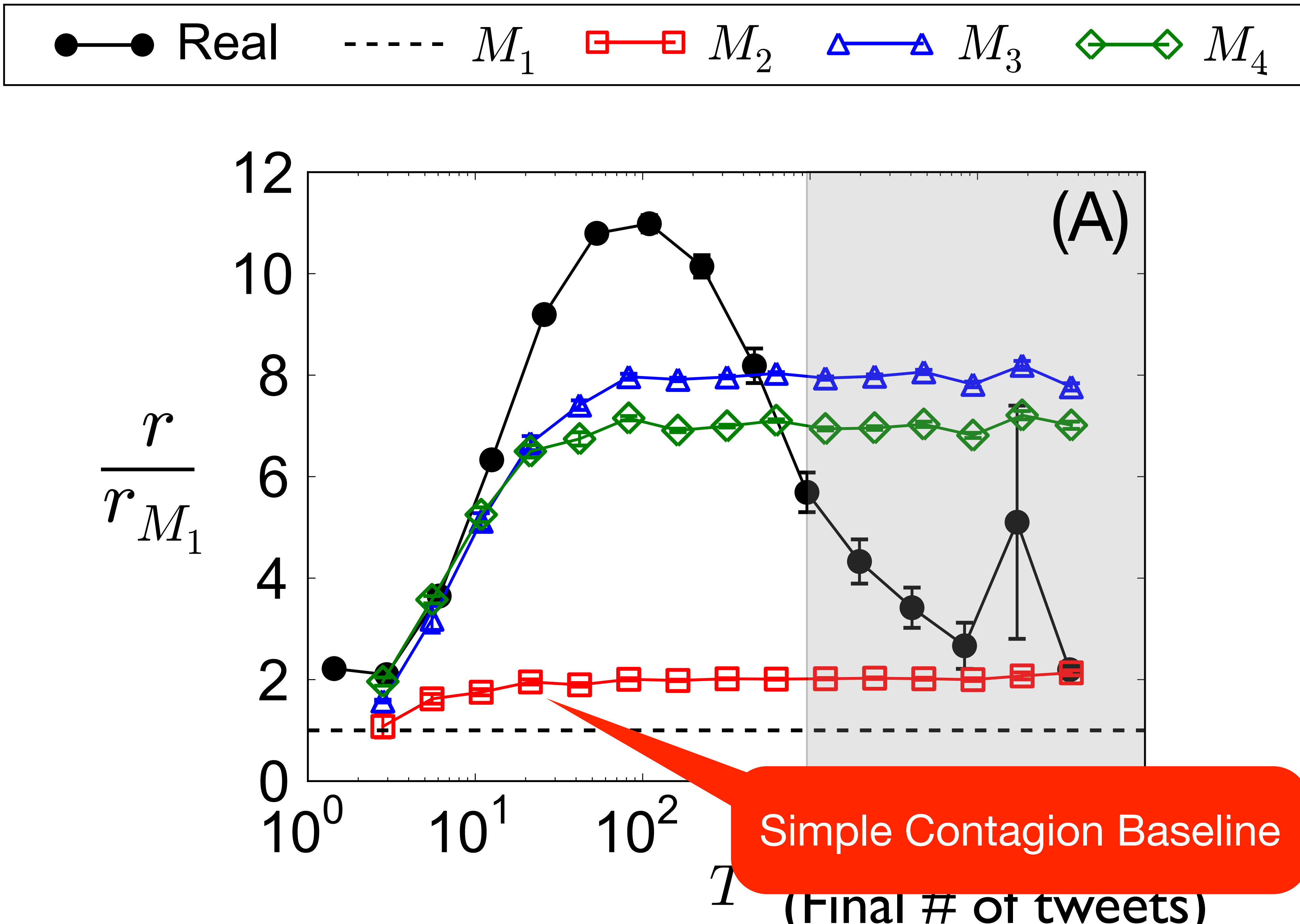
Normalize every one with M1

$r(h)/r_{M_1}(h)$ $H^t(h)/H^t_{M_1}(h)$

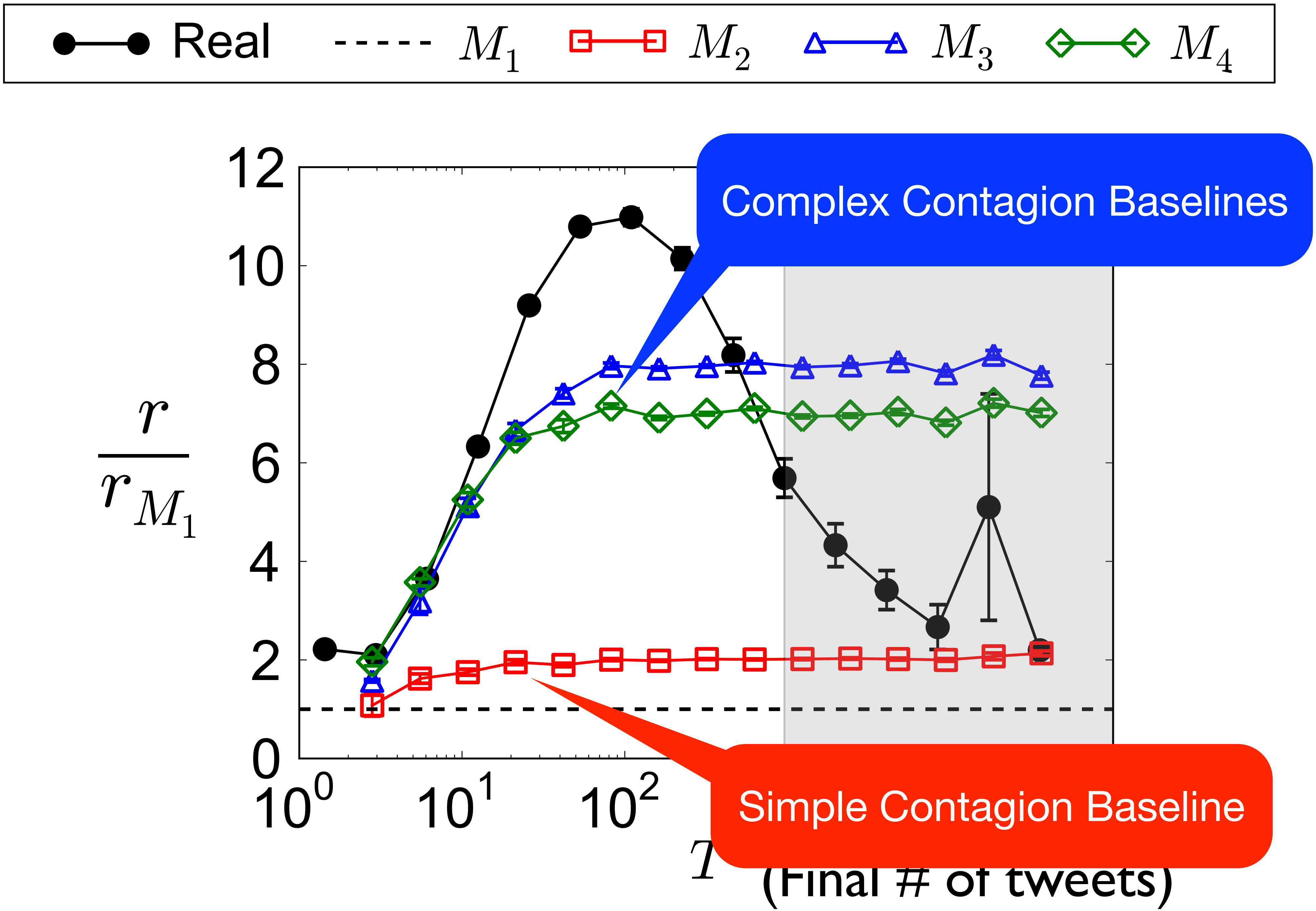
And use only the first 50 tweets.



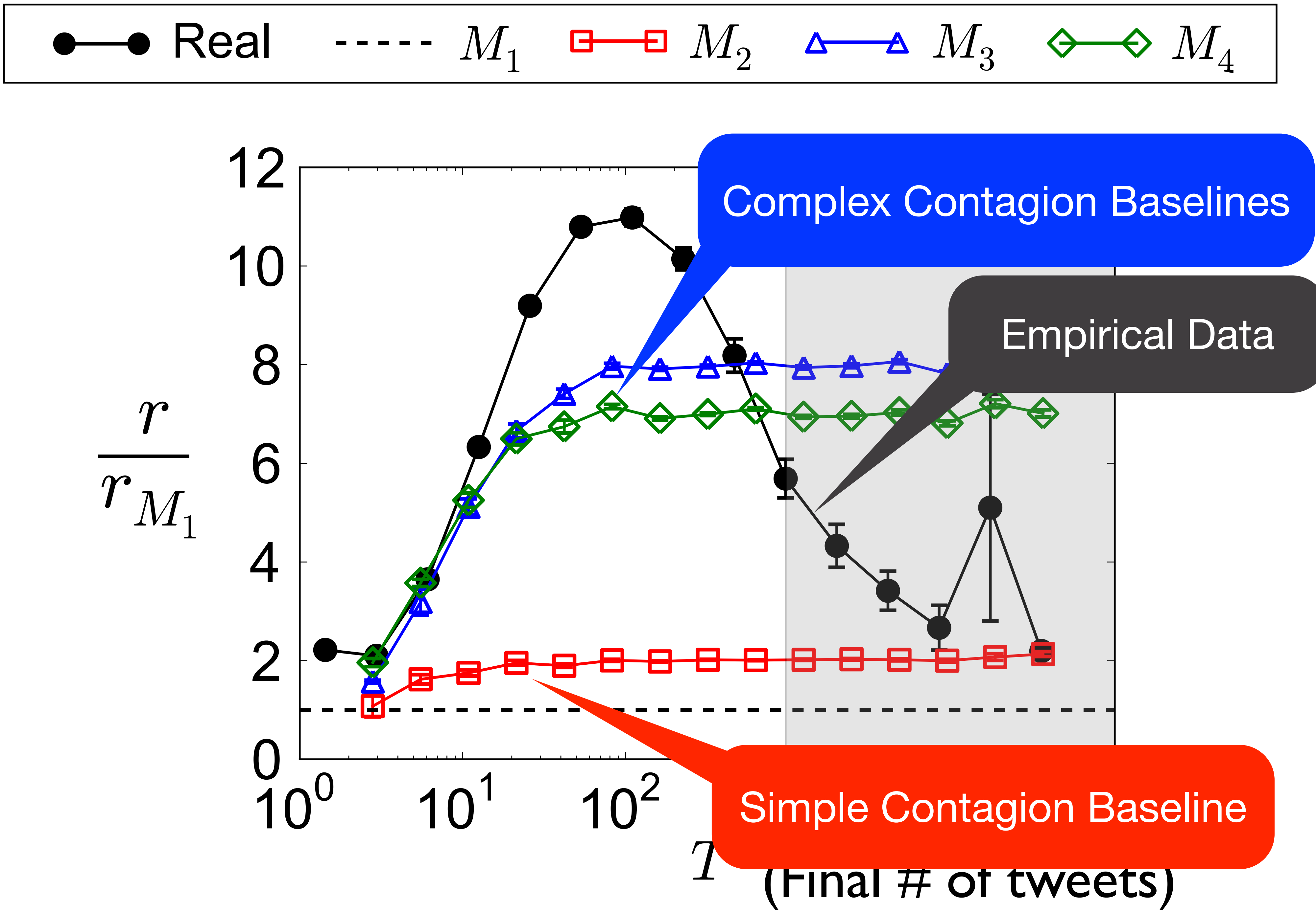
Real
$M_1$
$M_2$
$M_3$
$M_4$
(A)
$\frac{r}{r_{M_1}}$
0
2
4
6
8
10
12
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$
$T$ (Final # of tweets)



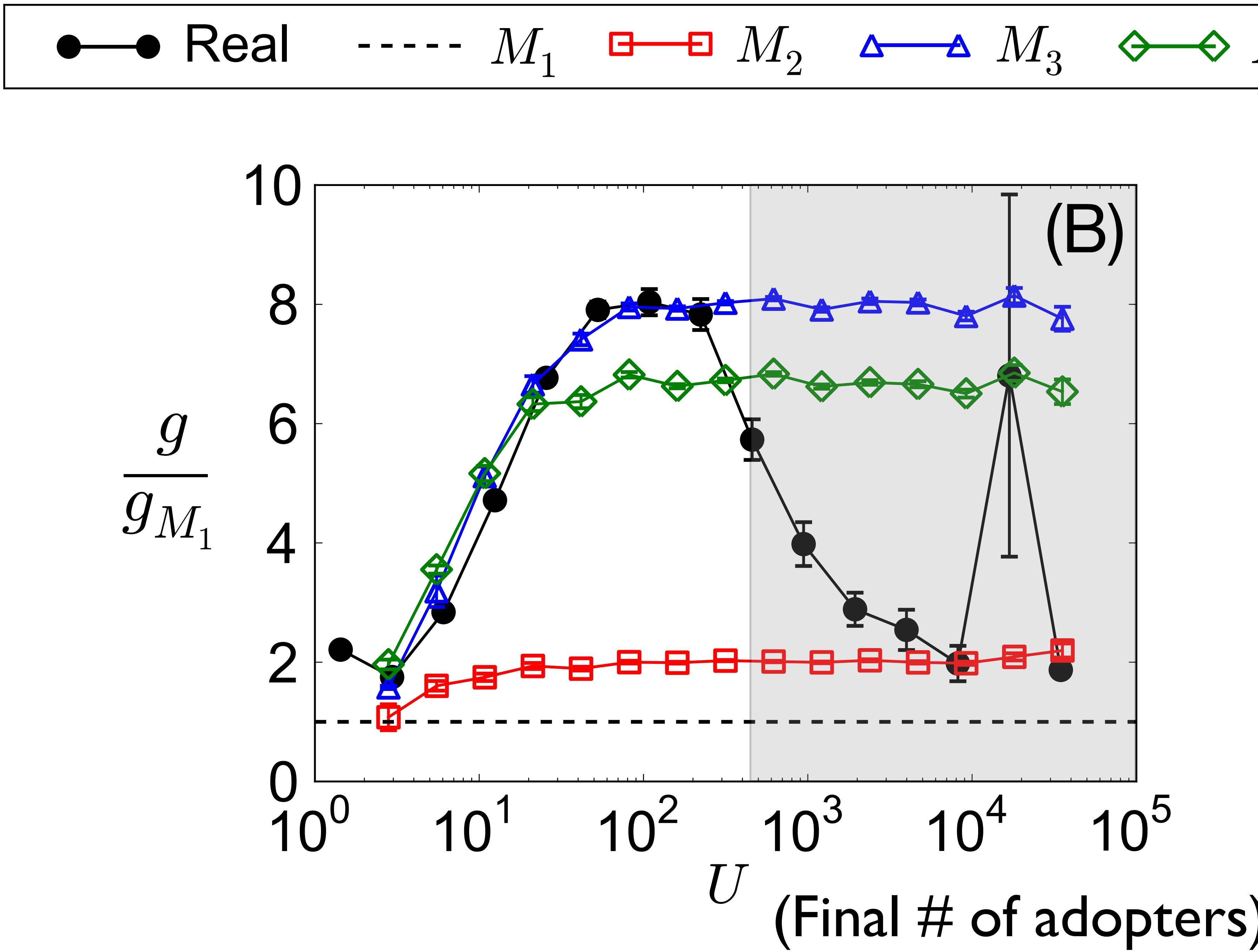
Real
$M_1$
$M_2$
$M_3$
$M_4$
(A)
$\frac{r}{r_{M_1}}$
12
10
8
6
4
2
0
$10^0$
$10^1$
$10^2$
$T$
(Final # of tweets)
Simple Contagion Baseline



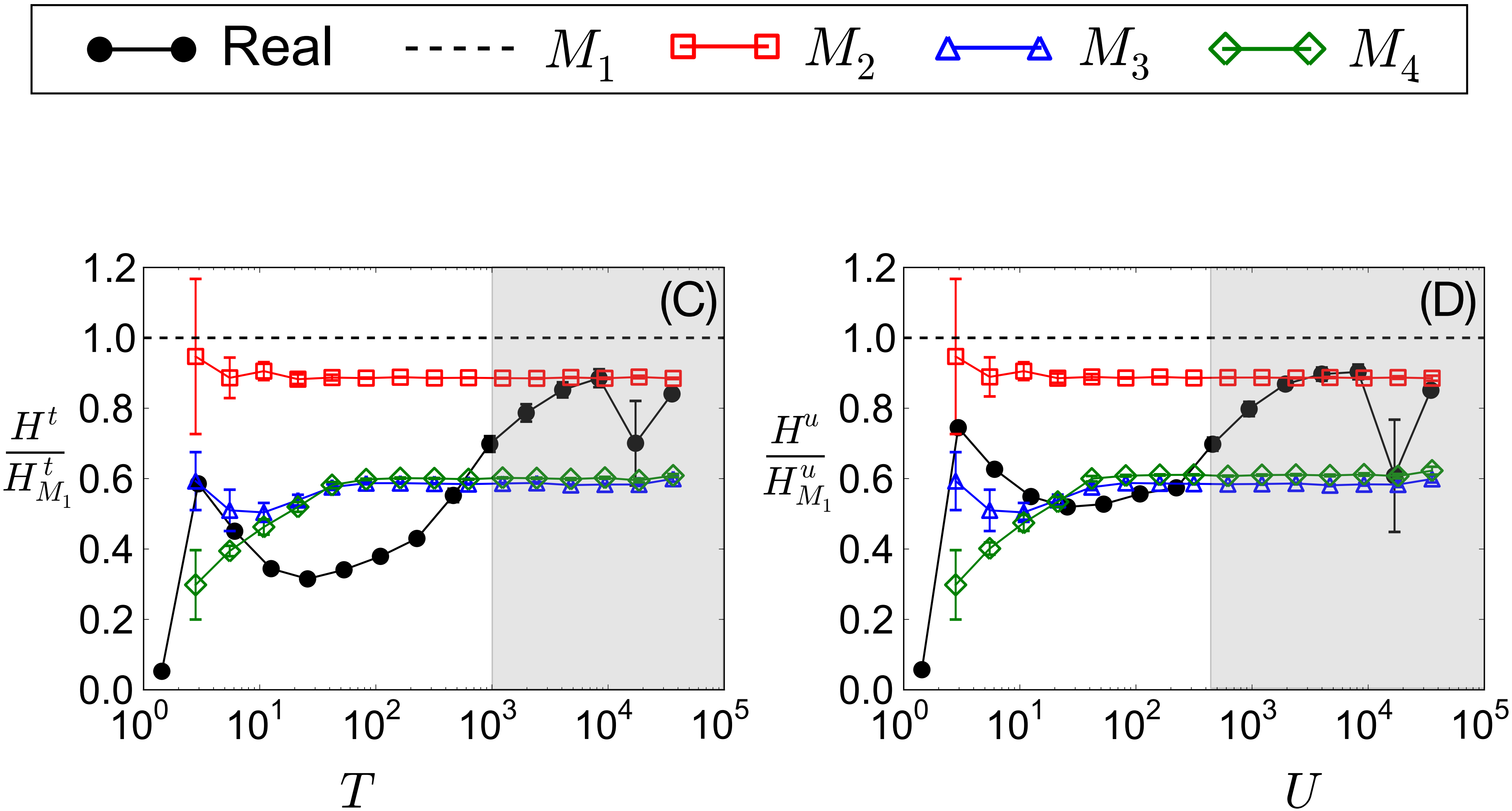
Real
$M_1$
$M_2$
$M_3$
$M_4$
Complex Contagion Baselines
Simple Contagion Baseline
$\frac{r}{r_{M_1}}$
12
10
8
6
4
2
0
$10^0$
$10^1$
$10^2$
$T$
(Final # of tweets)



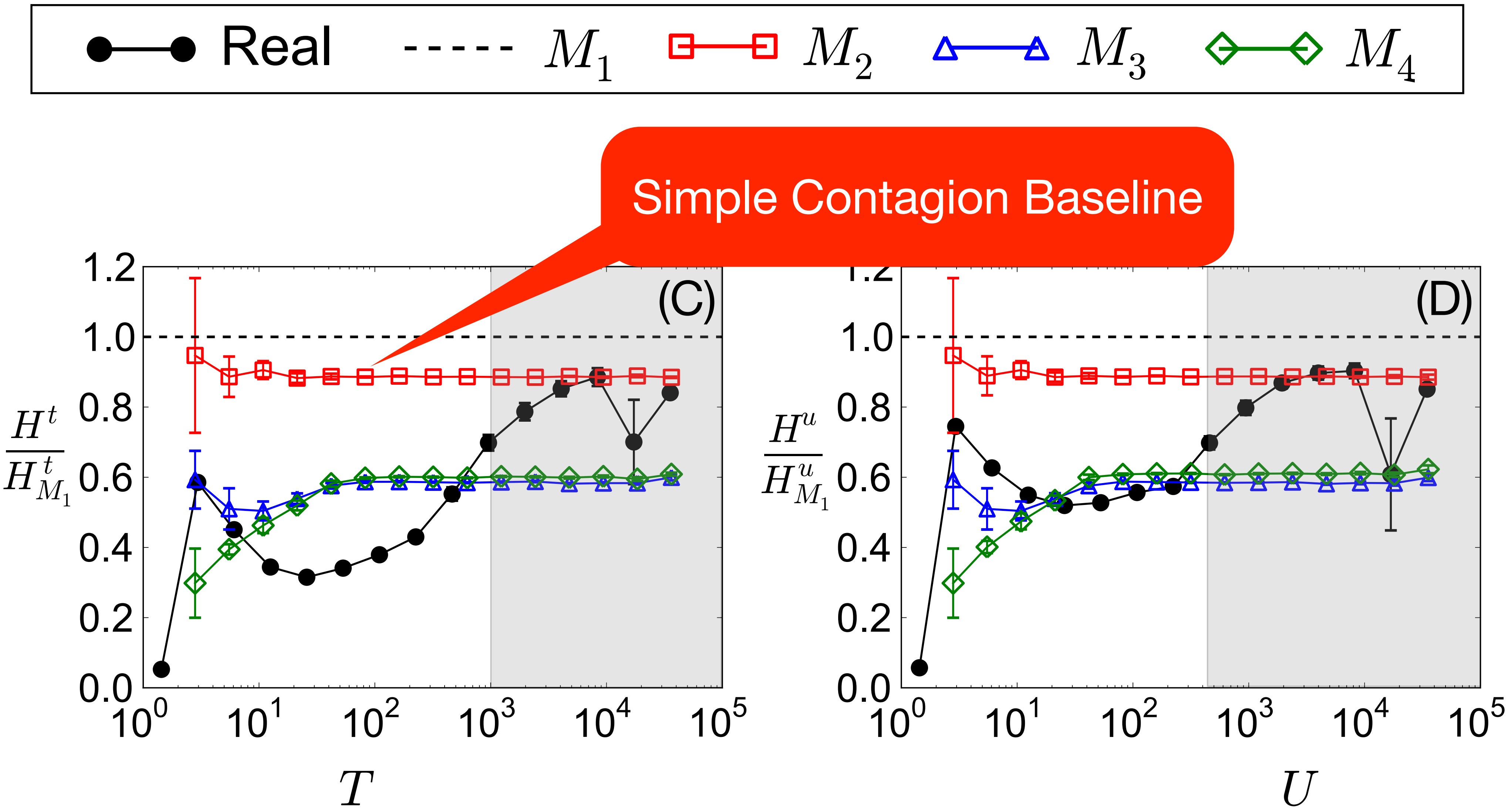
Real
$M_1$
$M_2$
$M_3$
$M_4$
Complex Contagion Baselines
Empirical Data
Simple Contagion Baseline
$\frac{r}{r_{M_1}}$
12
10
8
6
4
2
0
$10^0$
$10^1$
$10^2$
$T$
(Final # of tweets)



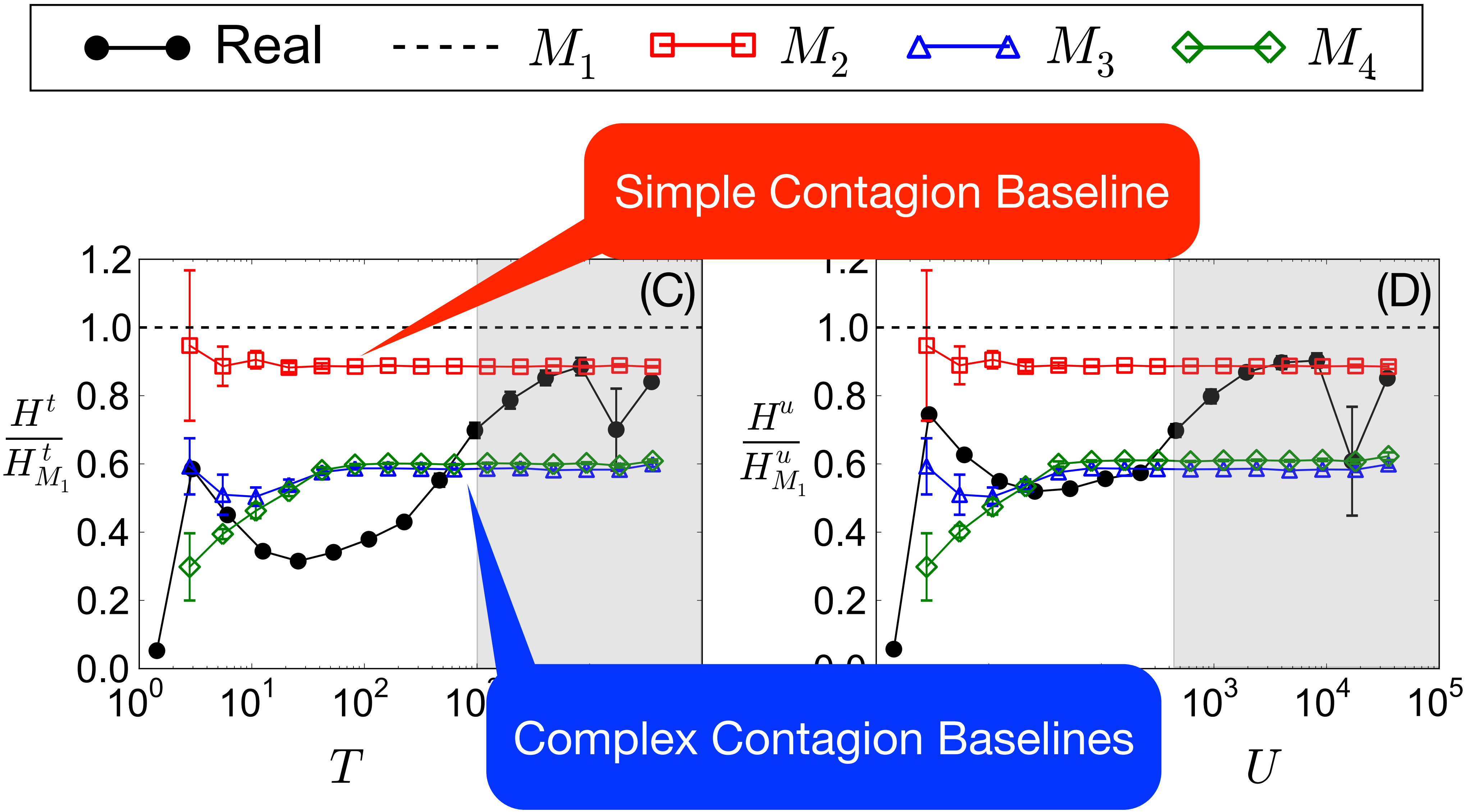
Real
$M_1$
$M_2$
$M_3$
$M_4$
(B)
$\frac{g}{g_{M_1}}$
0
2
4
6
8
10
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$
$U$
(Final # of adopters)



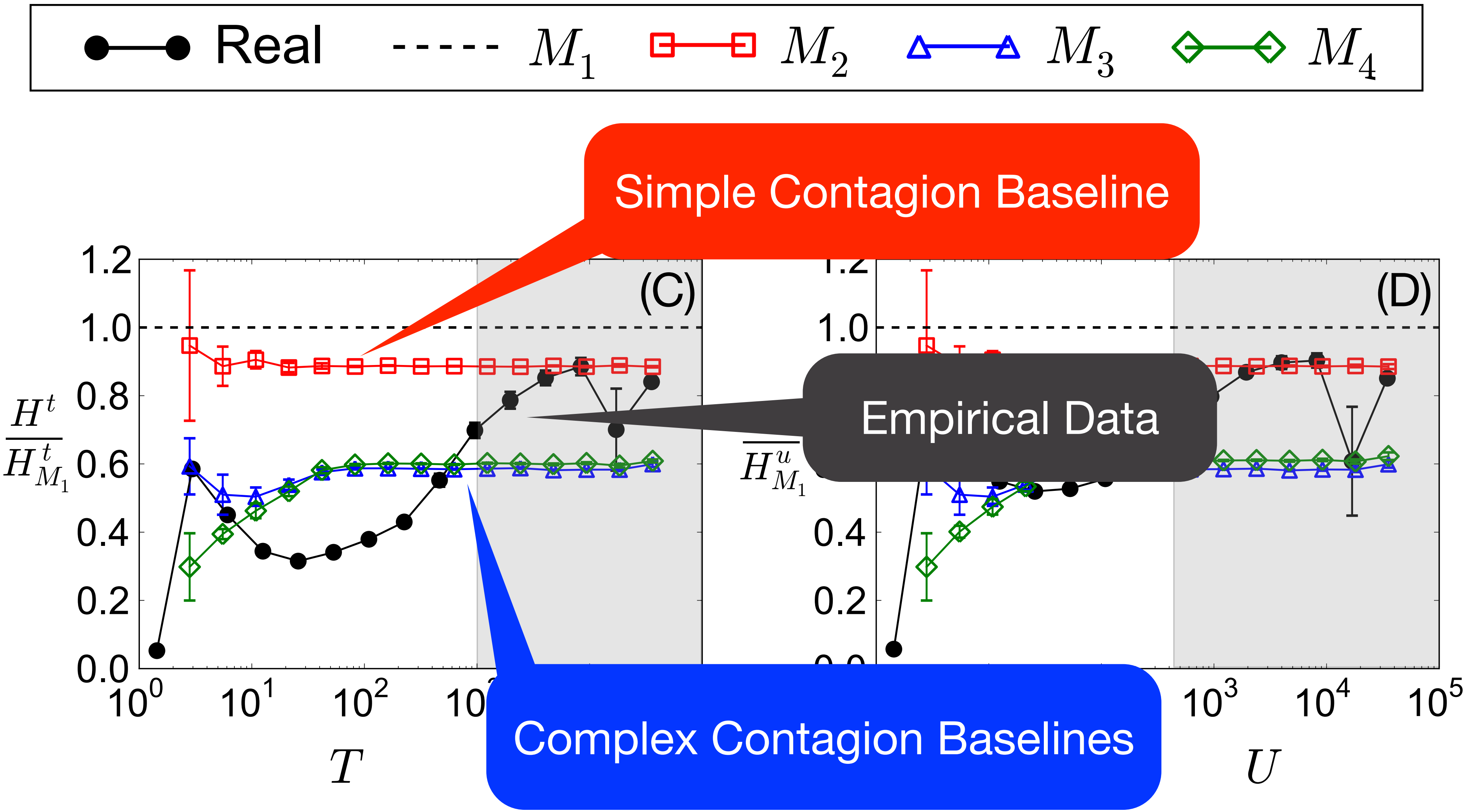
Real
$M_1$
$M_2$
$M_3$
$M_4$
(C)
(D)
$\frac{H^t}{H^t_{M_1}}$
$\frac{H^u}{H^u_{M_1}}$
$T$
$U$
1.2
1.0
0.8
0.6
0.4
0.2
0.0
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$



Real
$M_1$
$M_2$
$M_3$
$M_4$
Simple Contagion Baseline
(C)
(D)
$\frac{H^t}{H^t_{M_1}}$
$\frac{H^u}{H^u_{M_1}}$
$T$
$U$



Real
$M_1$
$M_2$
$M_3$
$M_4$
Simple Contagion Baseline
Complex Contagion Baselines
(C)
(D)
$\frac{H^t}{H^t_{M_1}}$
$\frac{H^u}{H^u_{M_1}}$
$T$
$U$



Real
$M_1$
$M_2$
$M_3$
$M_4$
Simple Contagion Baseline
Empirical Data
Complex Contagion Baselines
(C)
(D)
$\frac{H^t}{H^t_{M_1}}$
$H^u_{M_1}$
$T$
$U$
0.0
0.2
0.4
0.6
0.8
1.0
1.2
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$

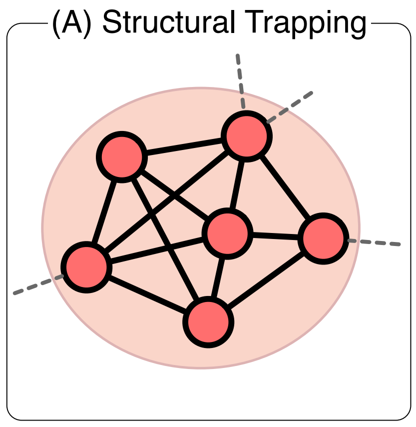
All memes are **not equal.**

**Unsuccessful** memes behave like **complex contagions**

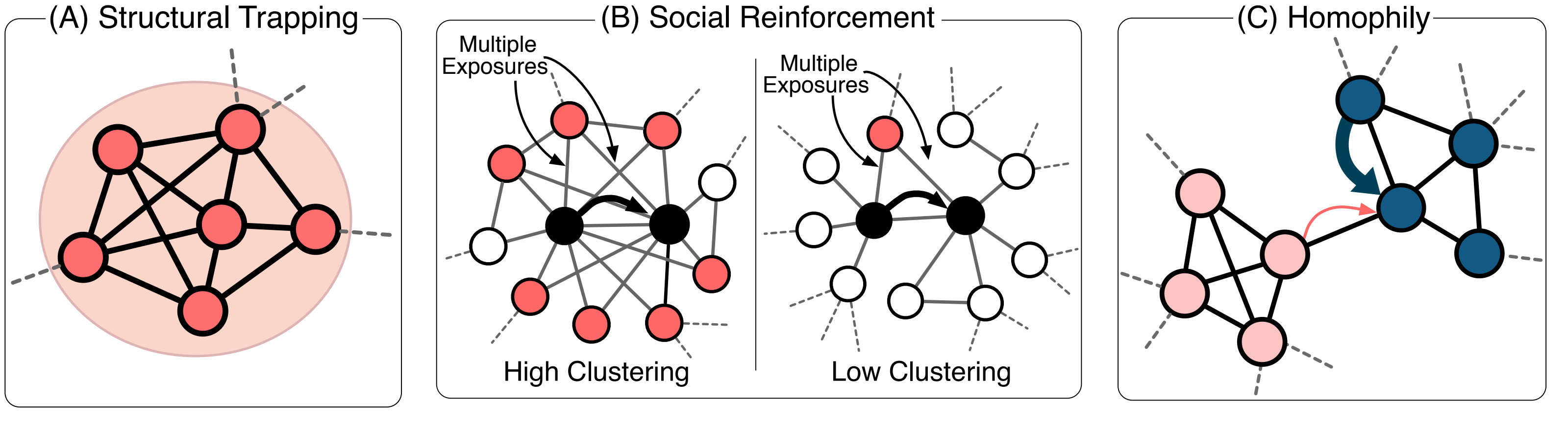
**Viral** memes behave like **simple contagions**

***Viral*** memes are literally ***viral.***

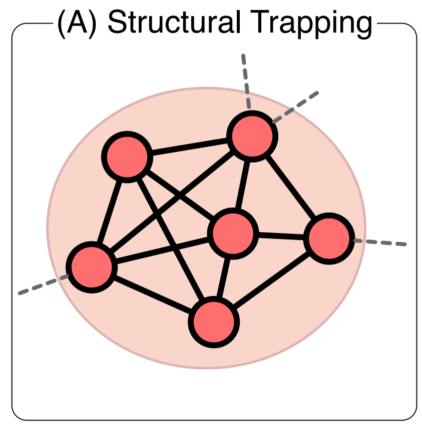
# Simple contagion



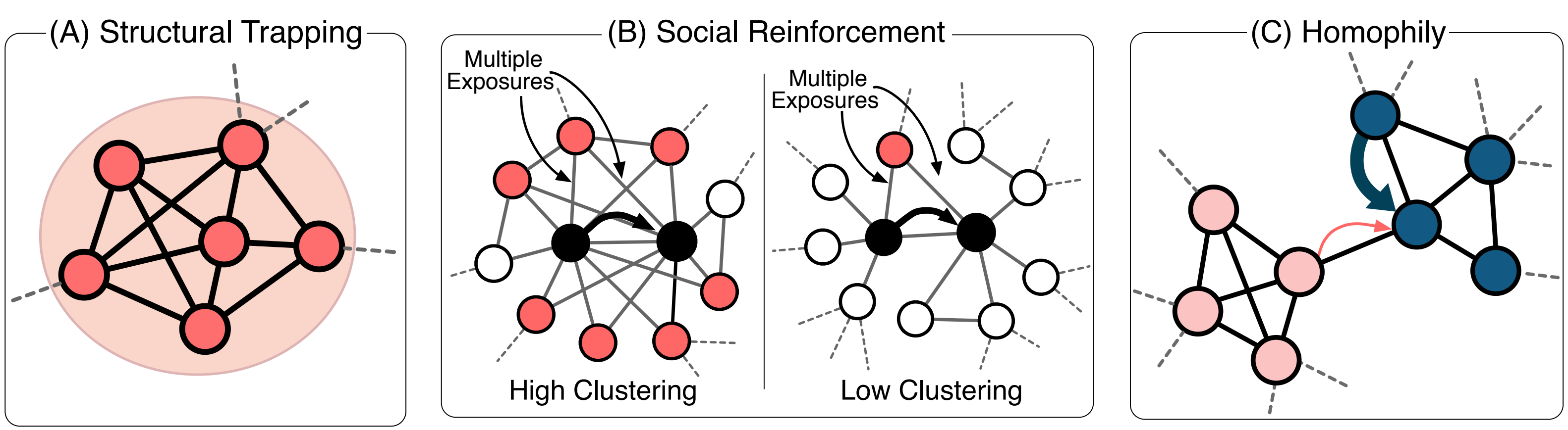

# Complex contagion



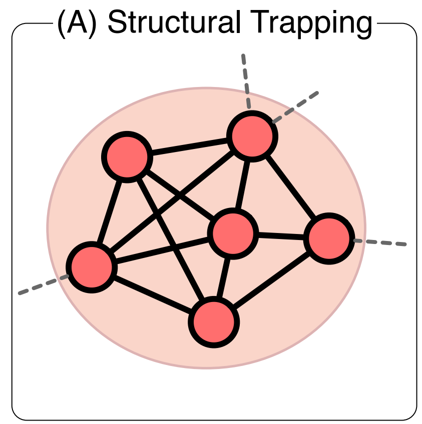
# Viral memes



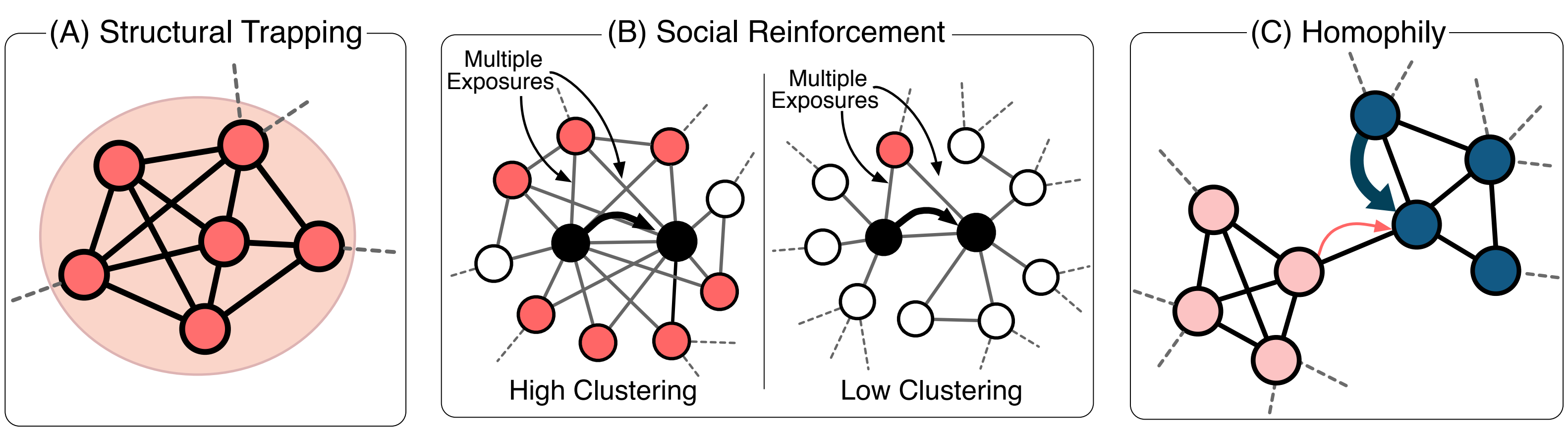

# Complex contagion



# Viral memes




# Non-viral memes



# Another perspective

Each community

~ interest group

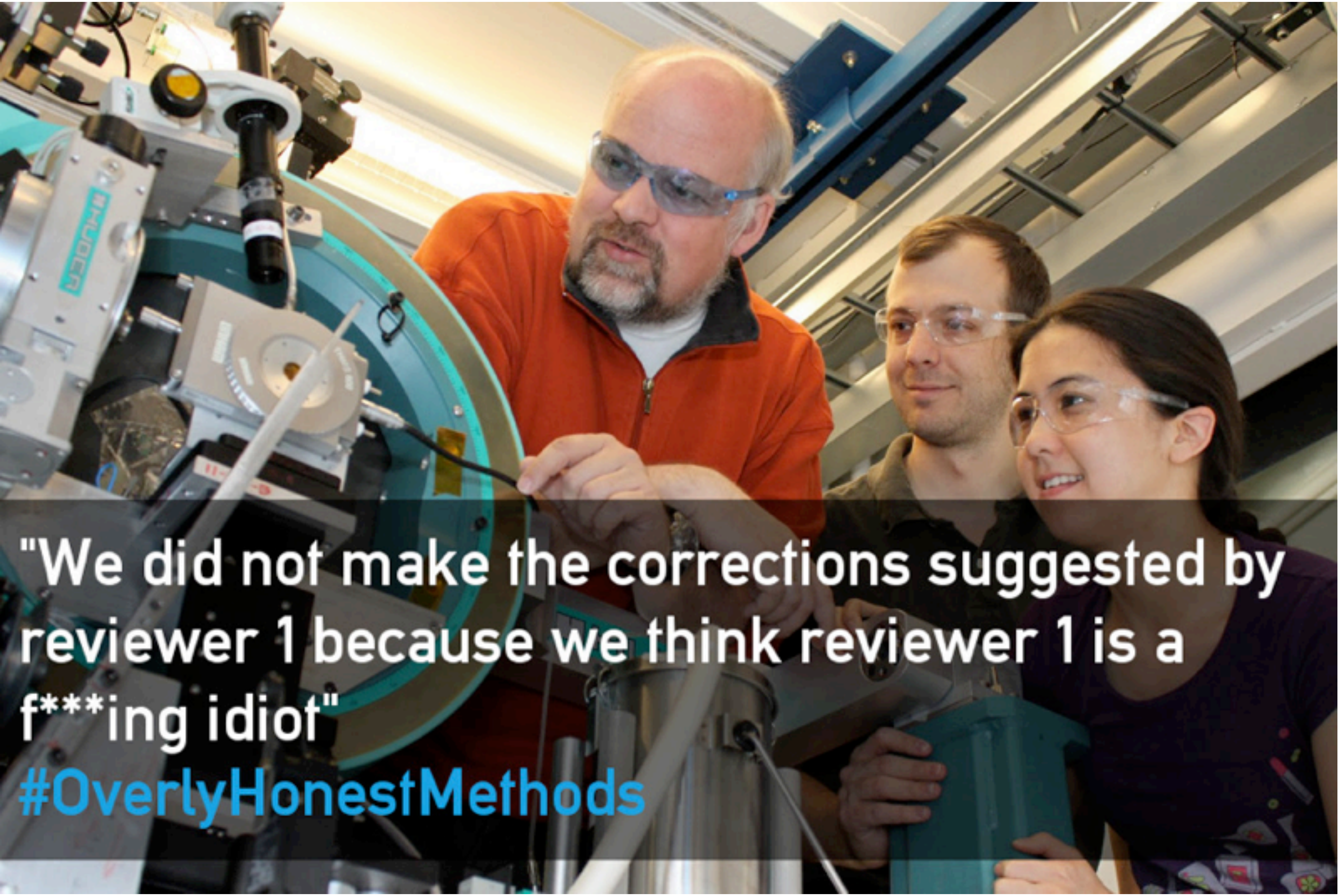
Concentrated in one community

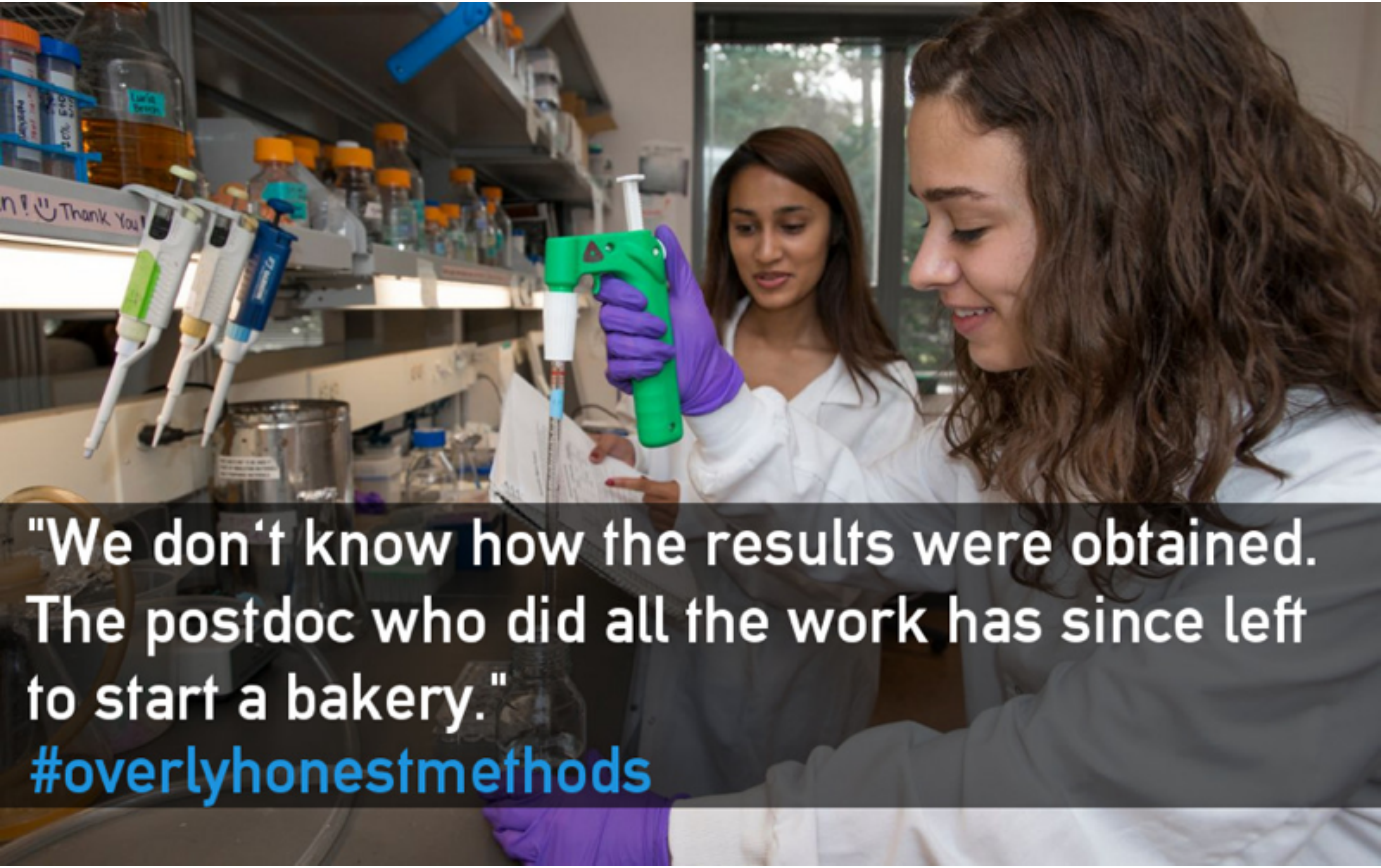
The meme only appeals to the population

Distributed throughout many communities

↓

The meme appeals to the general population


"We did not make the corrections suggested by reviewer 1 because we think reviewer 1 is a f***ing idiot"
#OverlyHonestMethods

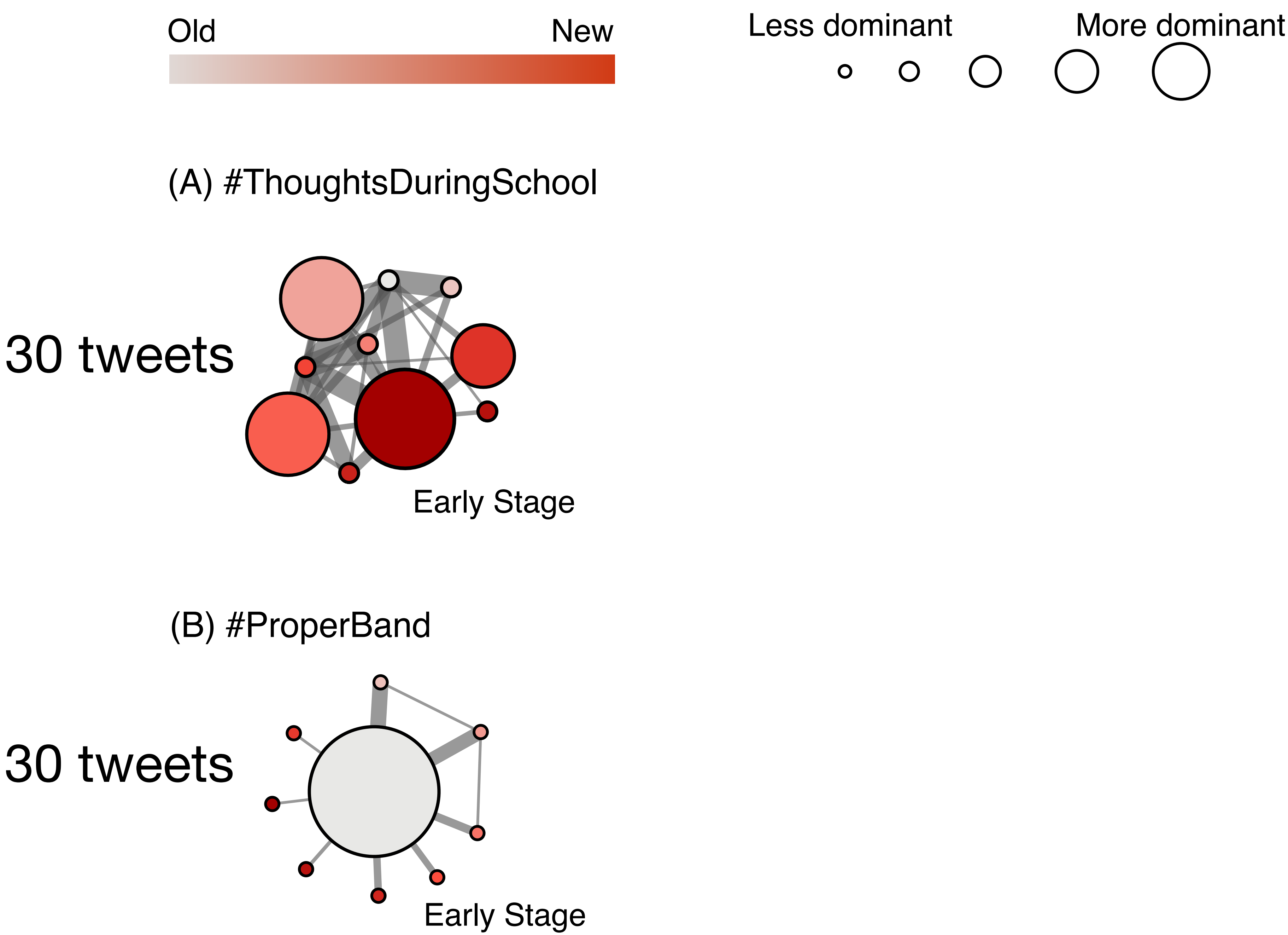

"We don't know how the results were obtained. The postdoc who did all the work has since left to start a bakery."
#overlyhonestmethods
Thank You!

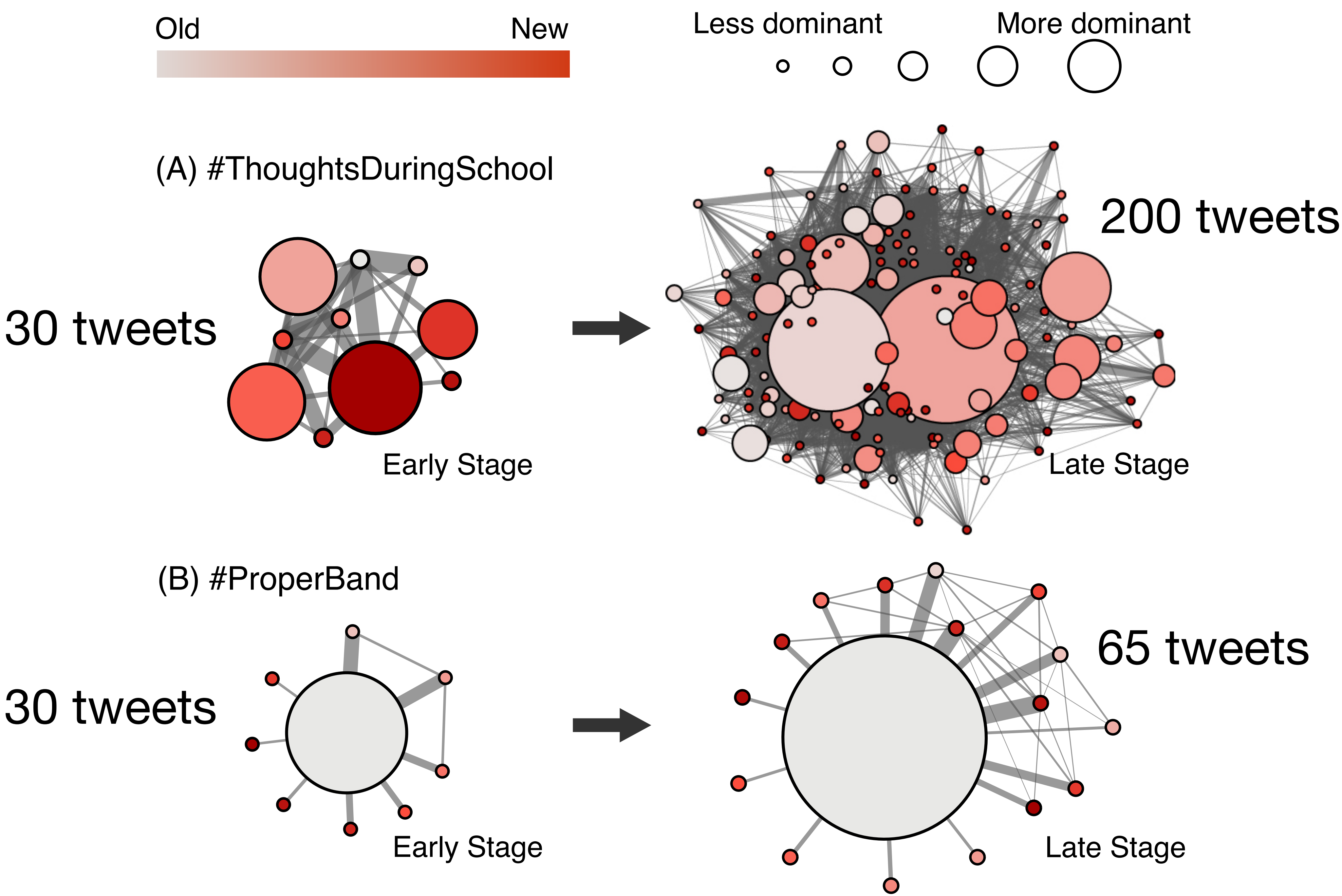
**Viral** memes are literally like **viruses**.

**Viral** memes are **attractive to everyone**.

Then, can we use this information to predict viral memes?

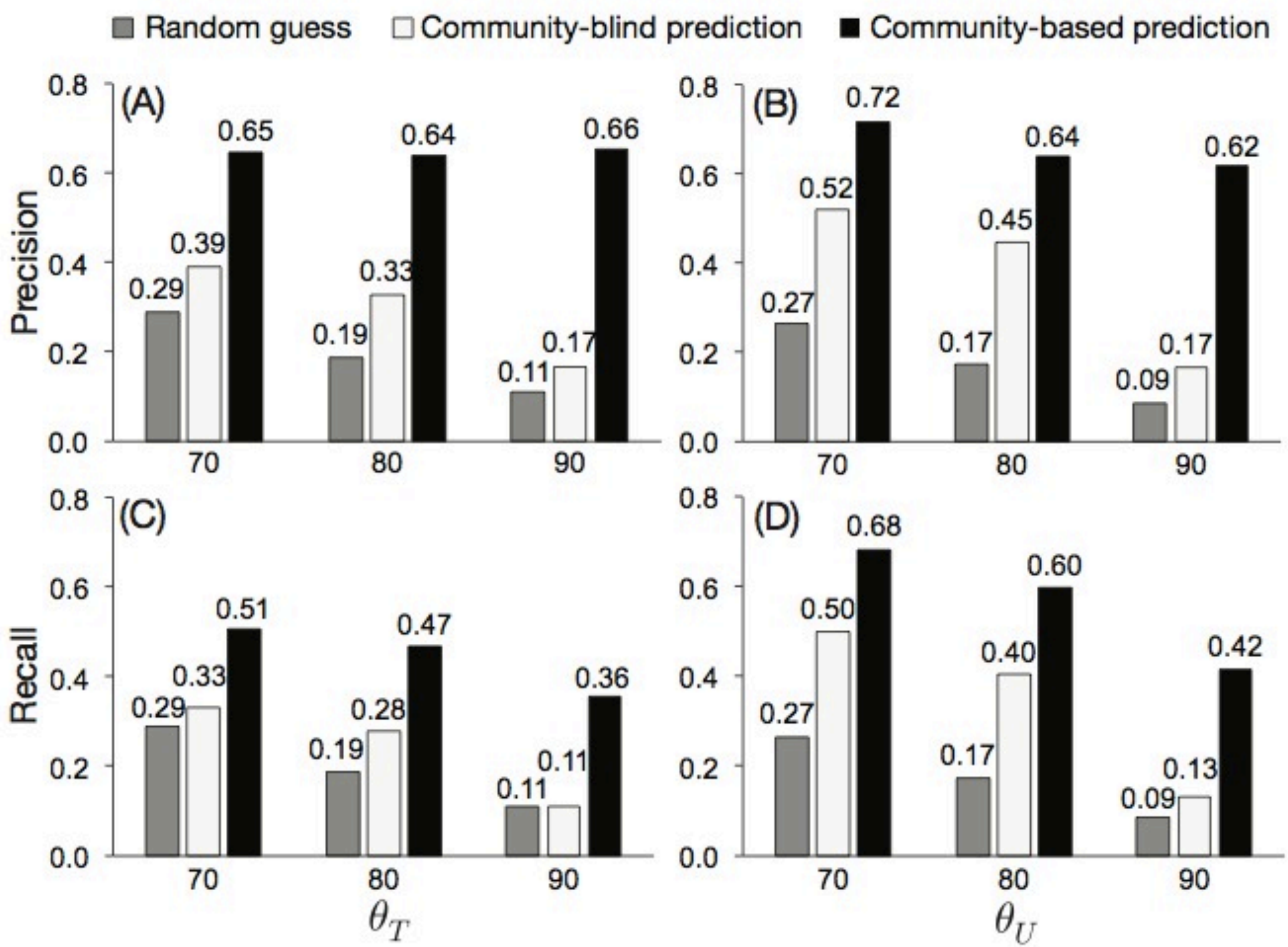
Task: Given the **network structure** and **early tweets**, predict the **final popularity**.

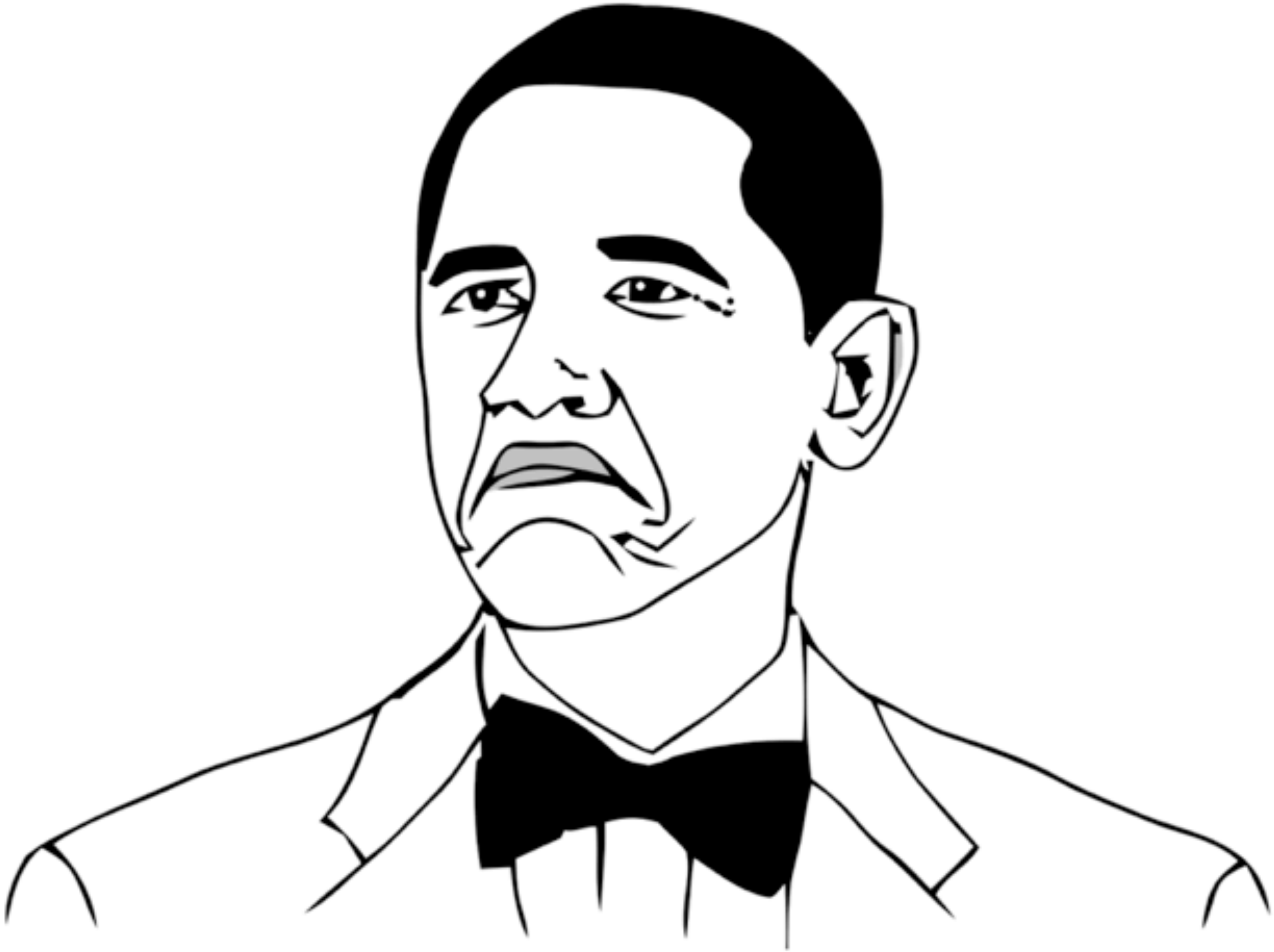

Old
New
Less dominant
More dominant
(A) #ThoughtsDuringSchool
30 tweets
Early Stage
(B) #ProperBand
30 tweets
Early Stage



Old
New
Less dominant
More dominant
(A) #ThoughtsDuringSchool
30 tweets
Early Stage
200 tweets
Late Stage
(B) #ProperBand
30 tweets
Early Stage
65 tweets
Late Stage

# With random forest classifier





NOT BAD

# Summary

- Communities give us invaluable information about spreading patterns of memes.
- We can predict viral memes by looking at communities
- Non-viral memes seems to be strongly affected by social reinforcement and homophily while viral memes are not.
- **Viral memes spread (literally) virally.**

Lilian Weng

Fil Menczer

**Virality Prediction and Community Structure in Social Networks**

[arxiv.org:1306.0158]