# Determining Factors of Information Spreading Processes

Dashun Wang

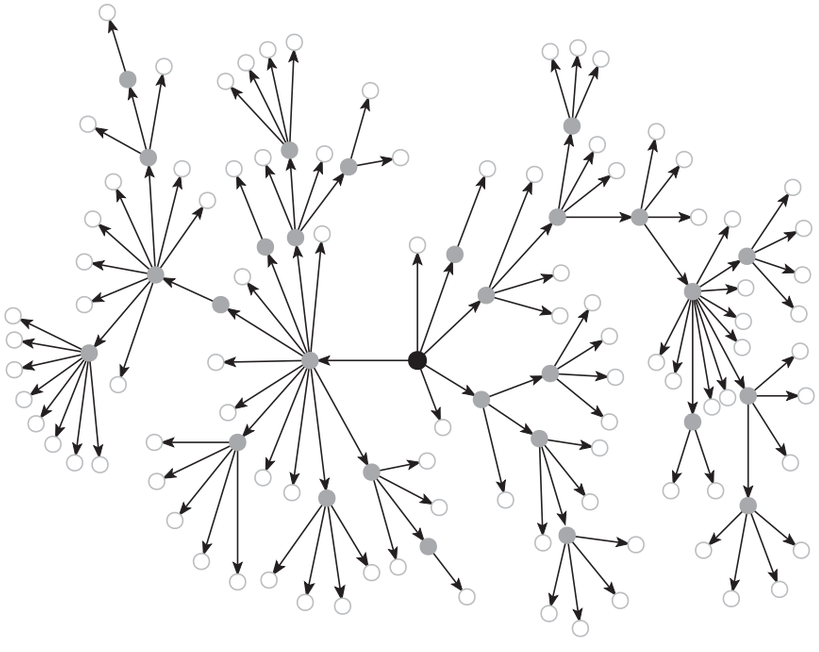
Northeastern

2011-06-07

HARVARD Medical School

# How a specific piece of information spreads?

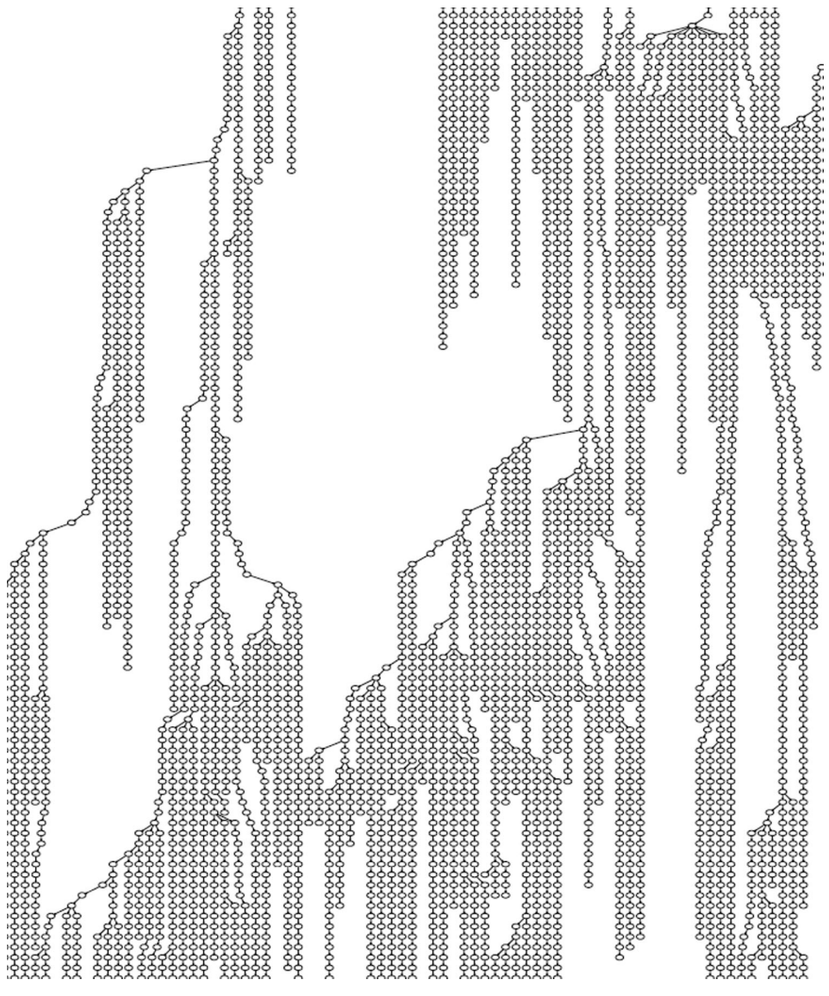
## Viral Marketing



José Luis Iribarren and E. Moro, *Phys. Rev. Lett.* 2009

D. Liben-Nowell and J. Kleinberg, *Proc Natl Acad Sci* 2008

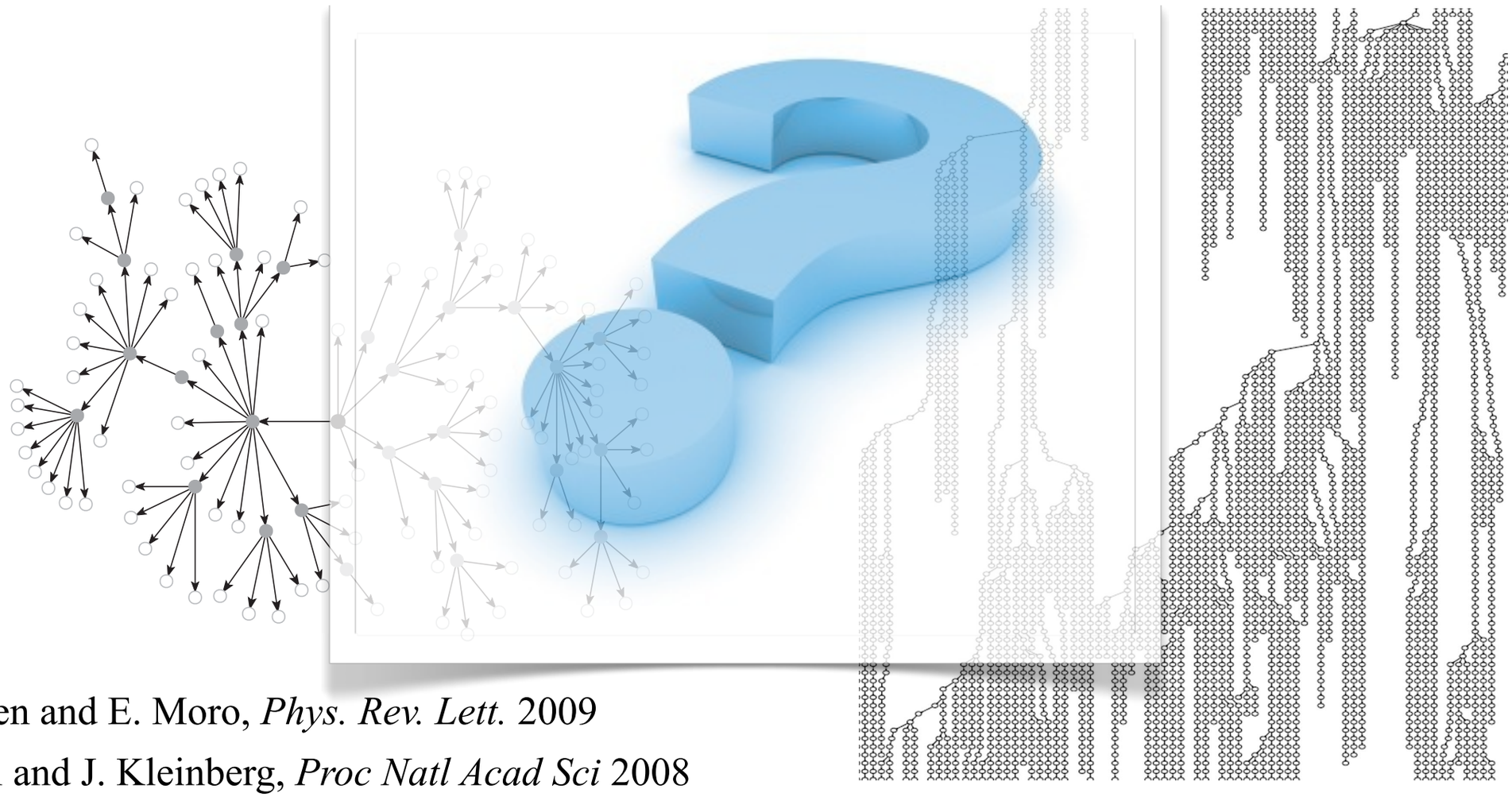
## Chain Letters



# What affects information spreading processes?

Viral Marketing

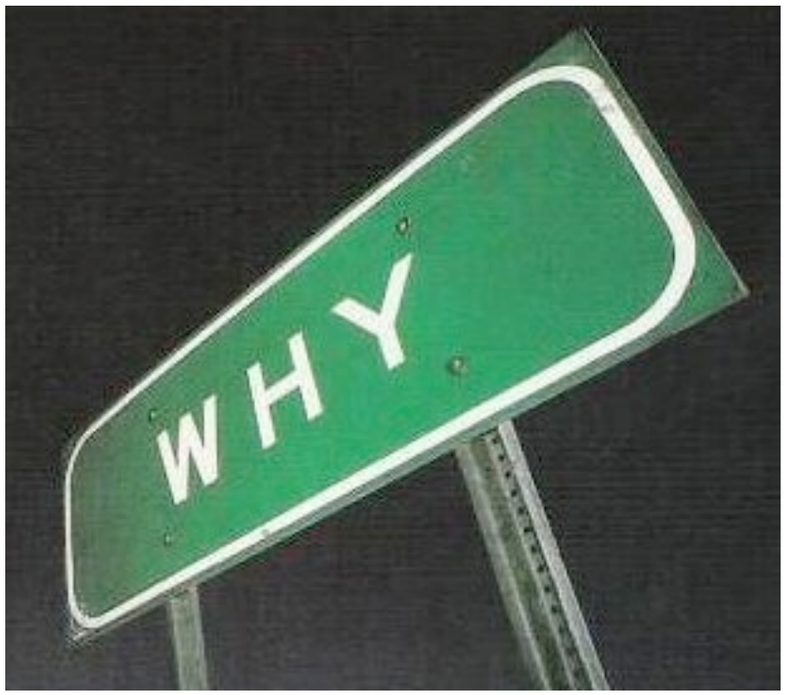
Chain Letters



José Luis Iribarren and E. Moro, *Phys. Rev. Lett.* 2009

D. Liben-Nowell and J. Kleinberg, *Proc Natl Acad Sci* 2008

# Information Spreading in Context




- Prediction models for information flow
- Assisting users to disseminate information more efficiently
- Protect digital information leakage
- Promote strategies to achieve expected coverage

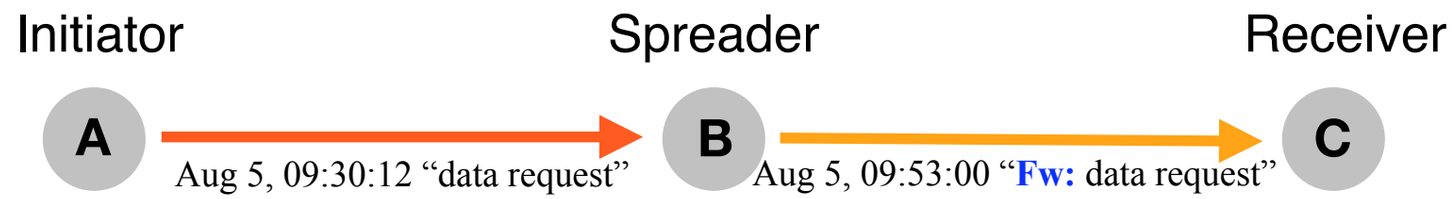
# Data

Detailed traces of social interactions

Forwarded Emails

↕

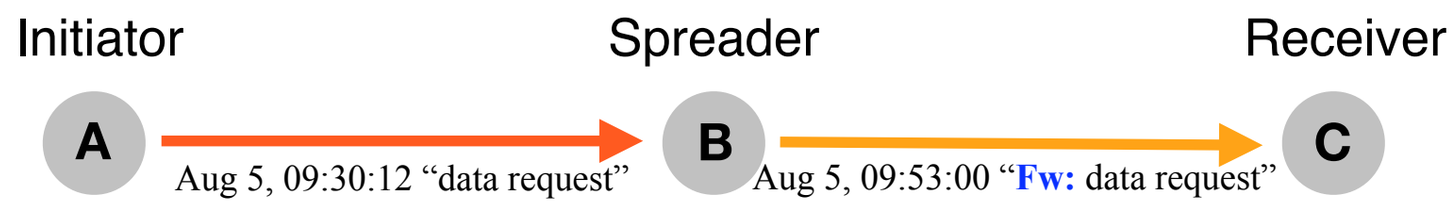
Spreading Processes



# Data

- Emails from 8000+ Employees — social network
- 2000+ Fw threads — ensemble of trees
- Information about the individuals — individual characteristics

  e.g.: performance, dept, job role
- Content of the emails — what the information is about

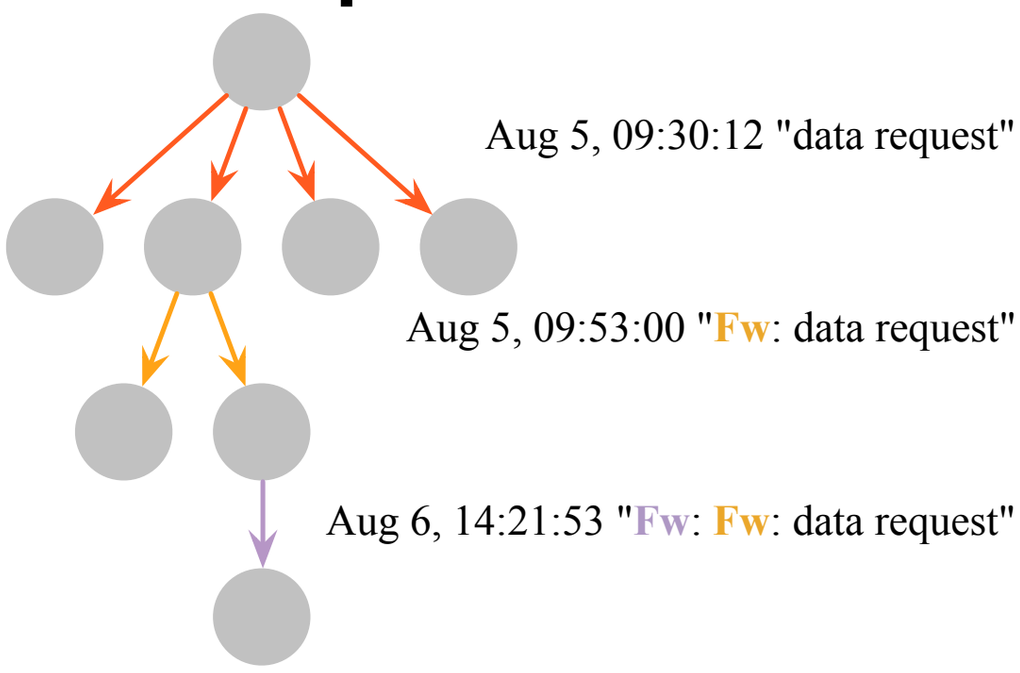


# Research Focus

## Microscopic Level

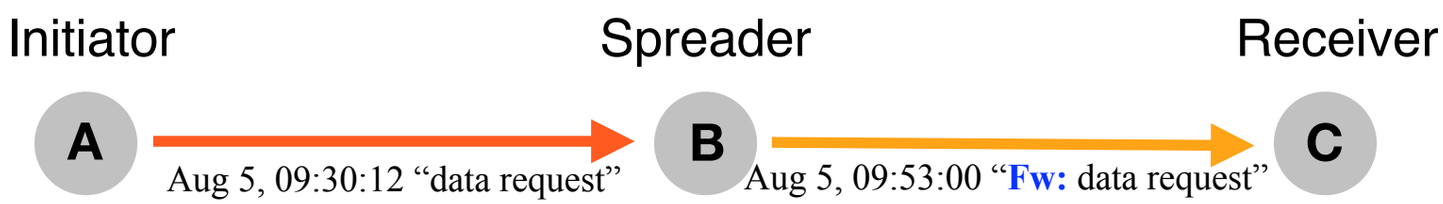

- To Whom one spreads the information
- Waiting time

## Macroscopic Level




- To how many people
- Overall coverage

# Microscopic Level



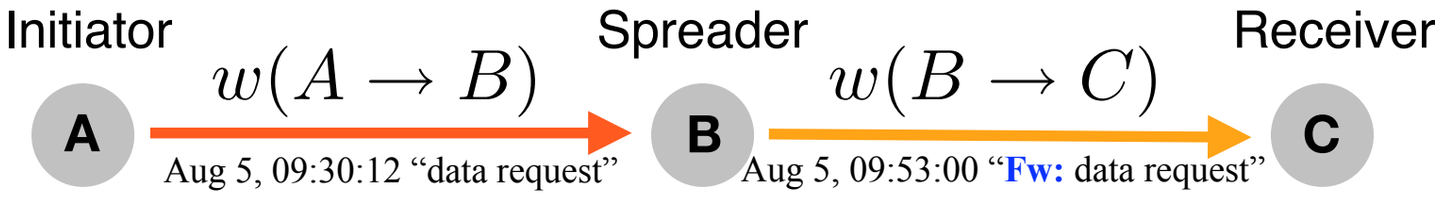

- The Underlying Social Network
- Information Content and Expertise
- Organizational Context
- Individual Characteristics

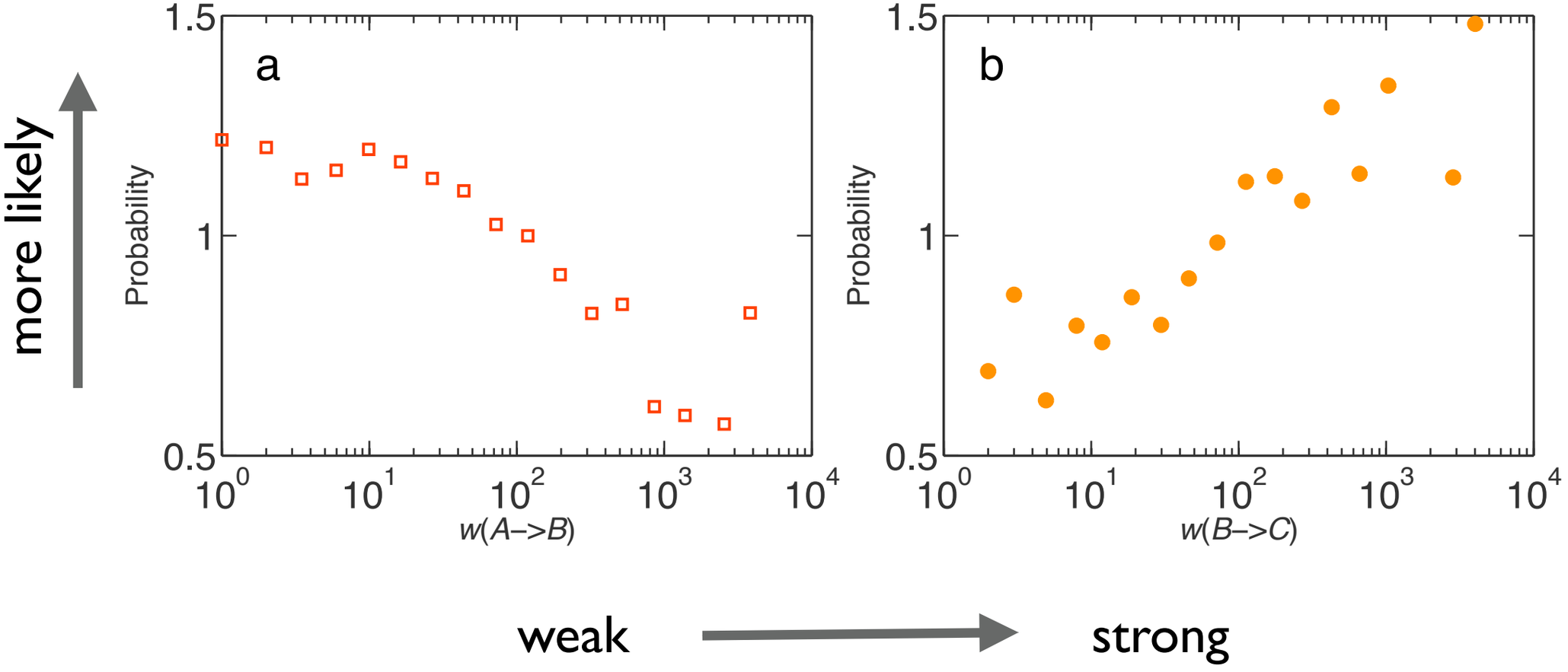
$P^{Fw}(q)$ Probability of having $q$ in forwarded emails

$P^{rand}(q)$ Same probability in normal emails

Probability Ratio: $P^{Fw}(q)/P^{rand}(q) \longleftrightarrow 1$

# The Underlying Social Network



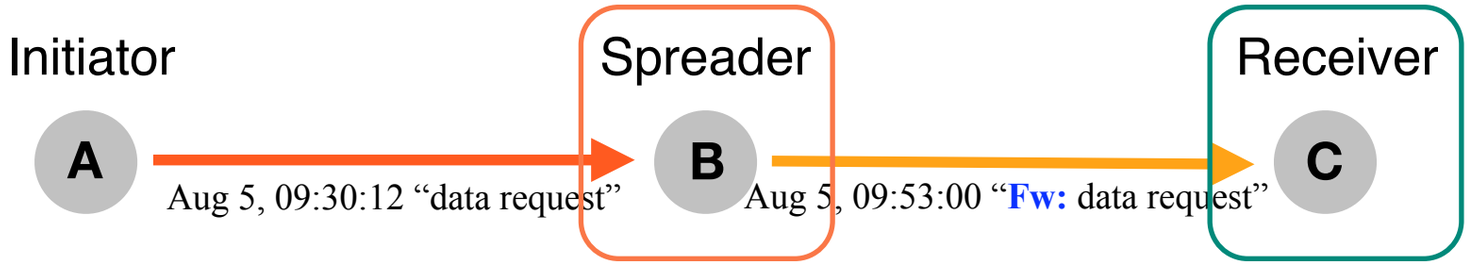



# Information Content and Expertise



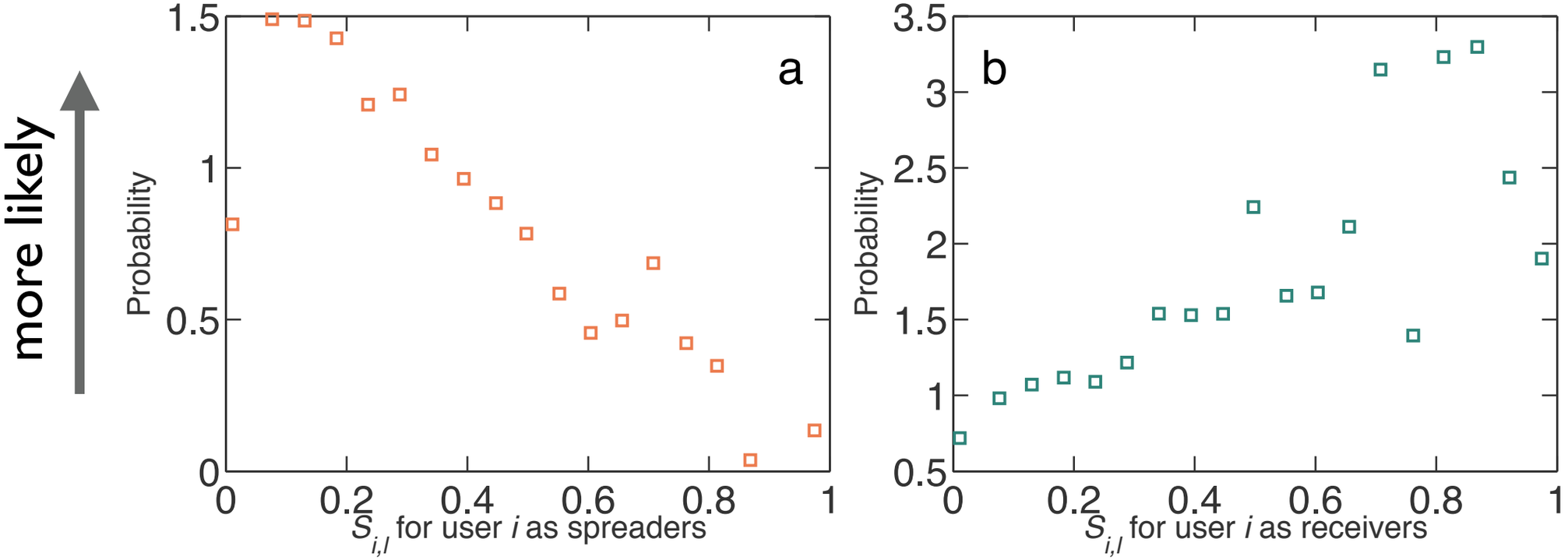
$$\mathcal{S}_{i,l} = \vec{v}_i \cdot \vec{v}_l / (\|\vec{v}_i\| \, \|\vec{v}_l\|)$$

# Information Content and Expertise

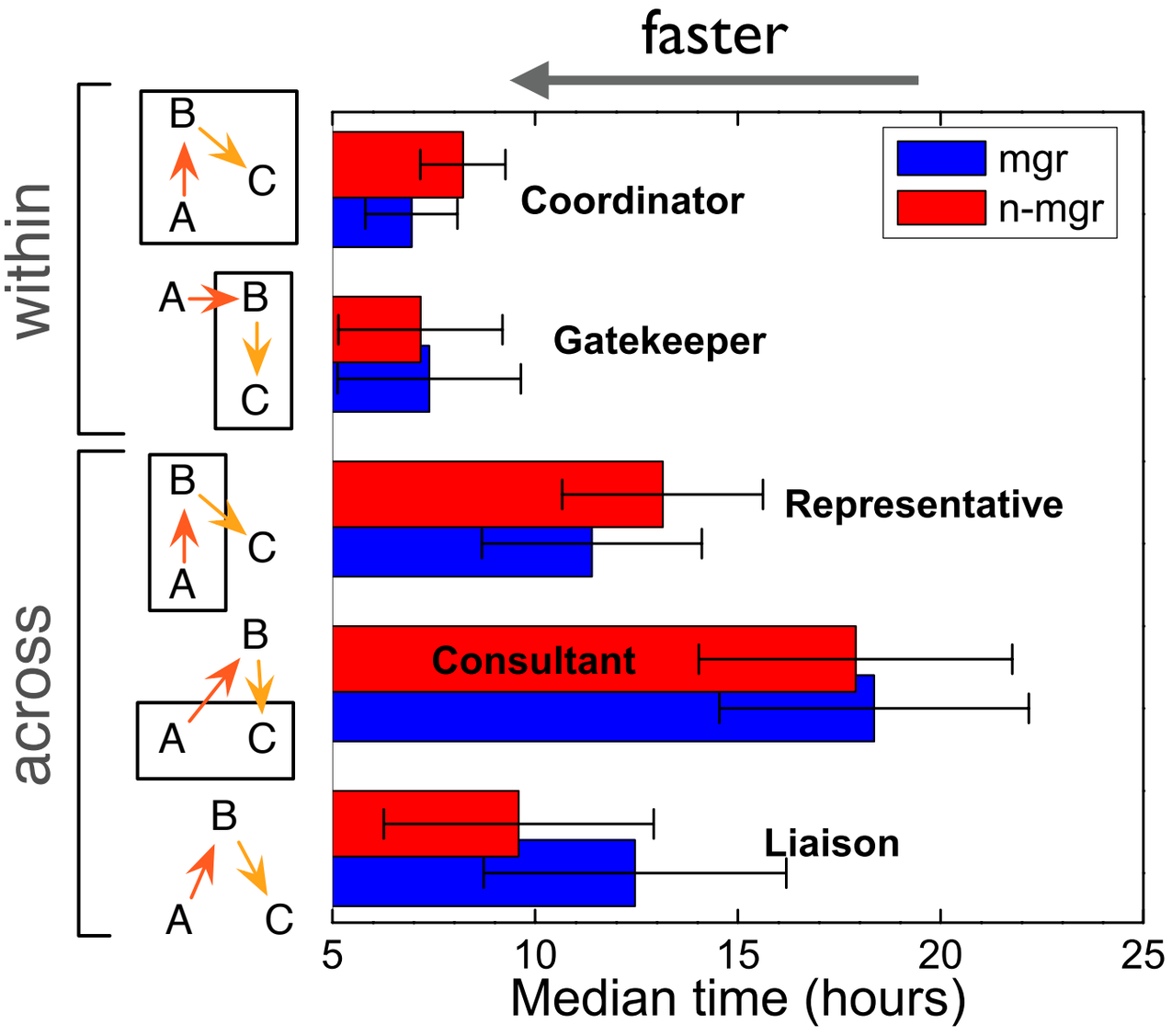




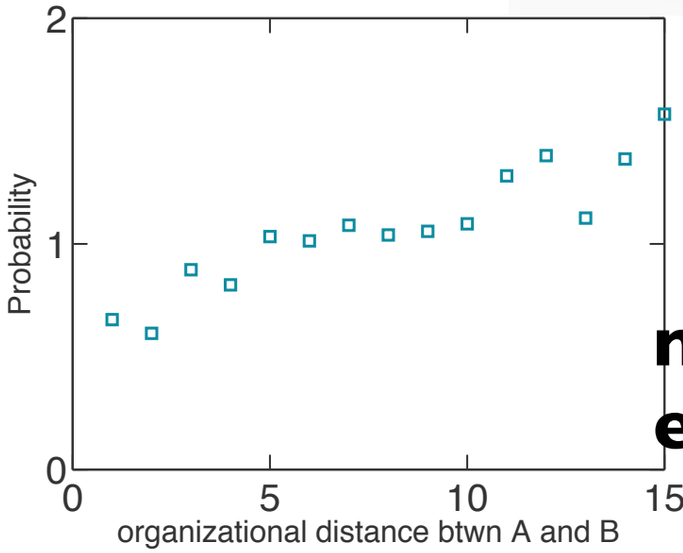
$$\mathcal{S}_{i,l} = \vec{v}_i \cdot \vec{v}_l / (\|\vec{v}_i\| \, \|\vec{v}_l\|)$$



# Organizational Context

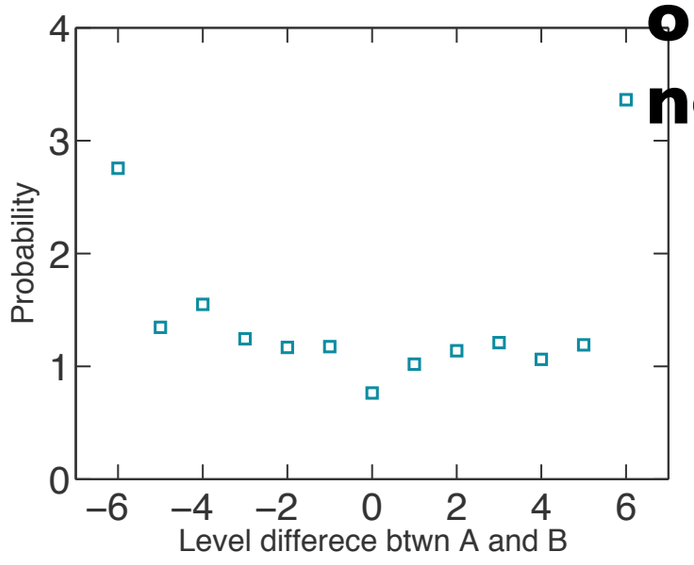



**social bottleneck**:
information
experiences delay
in inter-dept flows



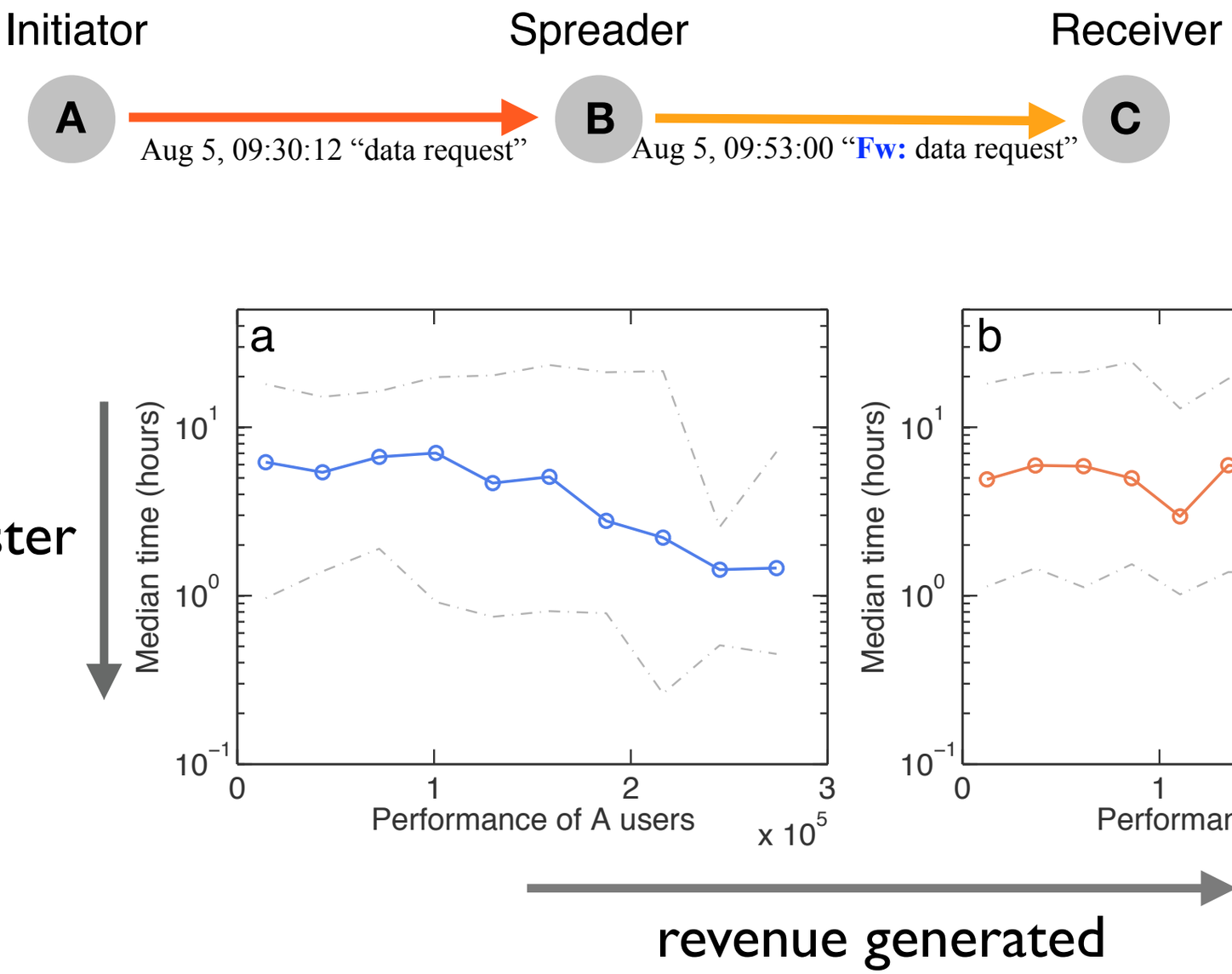

**non-homophily effect**

**on "formal" network**



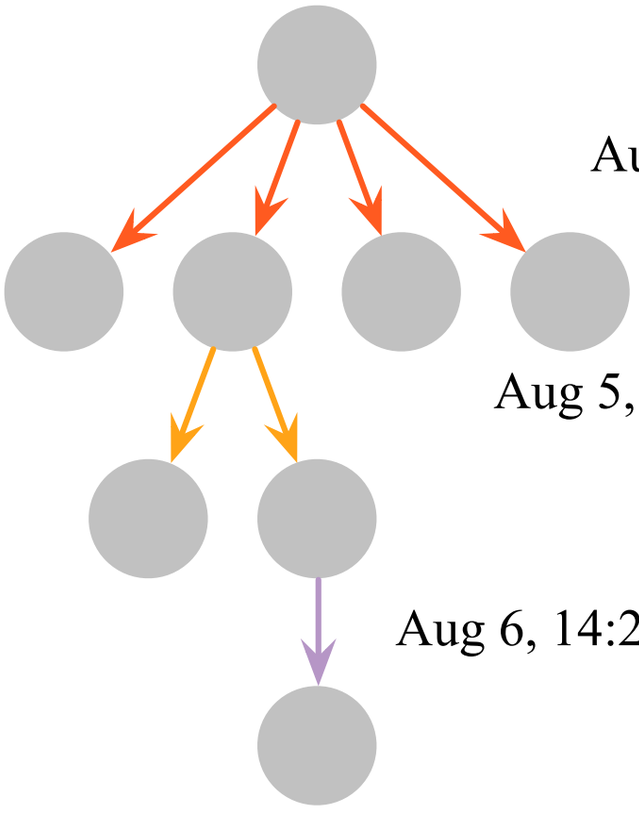

R. Gould and R. Fernandez. Structures of mediation: A formal approach to brokerage in transaction networks. *Sociological Methodology*, 19(1989):89–126, 1989.

# Individual Characterisitcs




The information waiting time appears to be const. for both initiators and spreaders, independent of individual performance

# Macroscopic Level

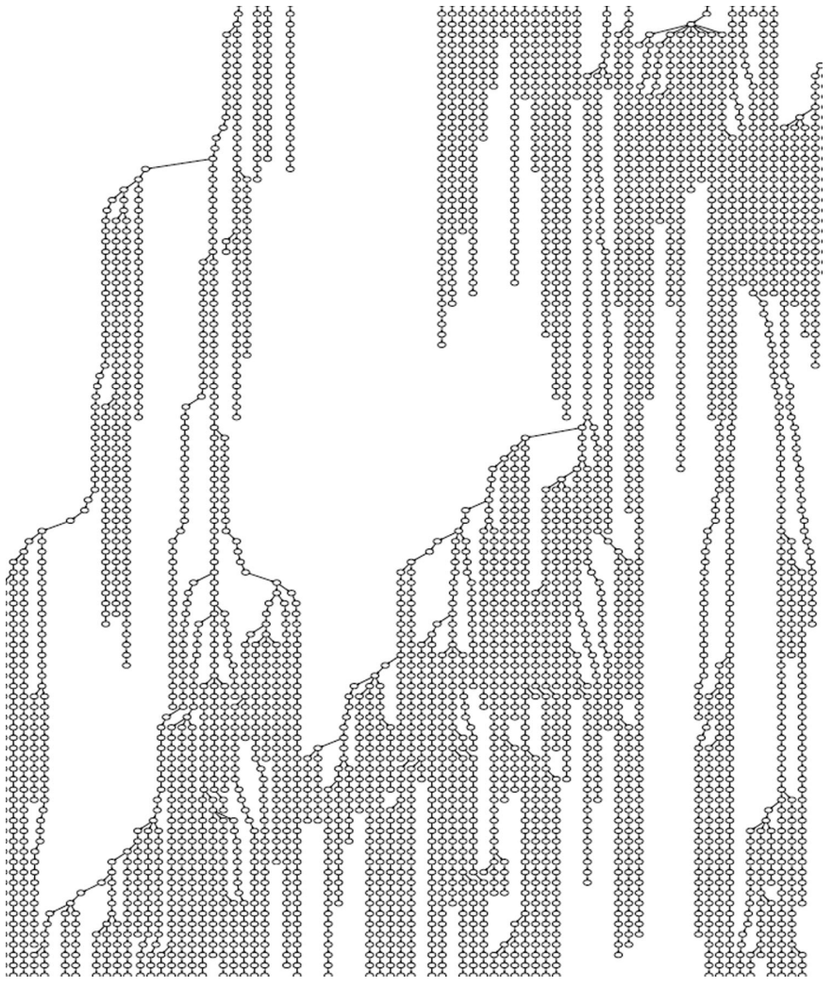



- Structural Properties of the spreading processes
- What's the best model for the observed structures

1. To how many people one would forward the information
2. What is the overall coverage

Is there a model that could fit the structures of the spreading process?

## Galton-Watson Branching Process



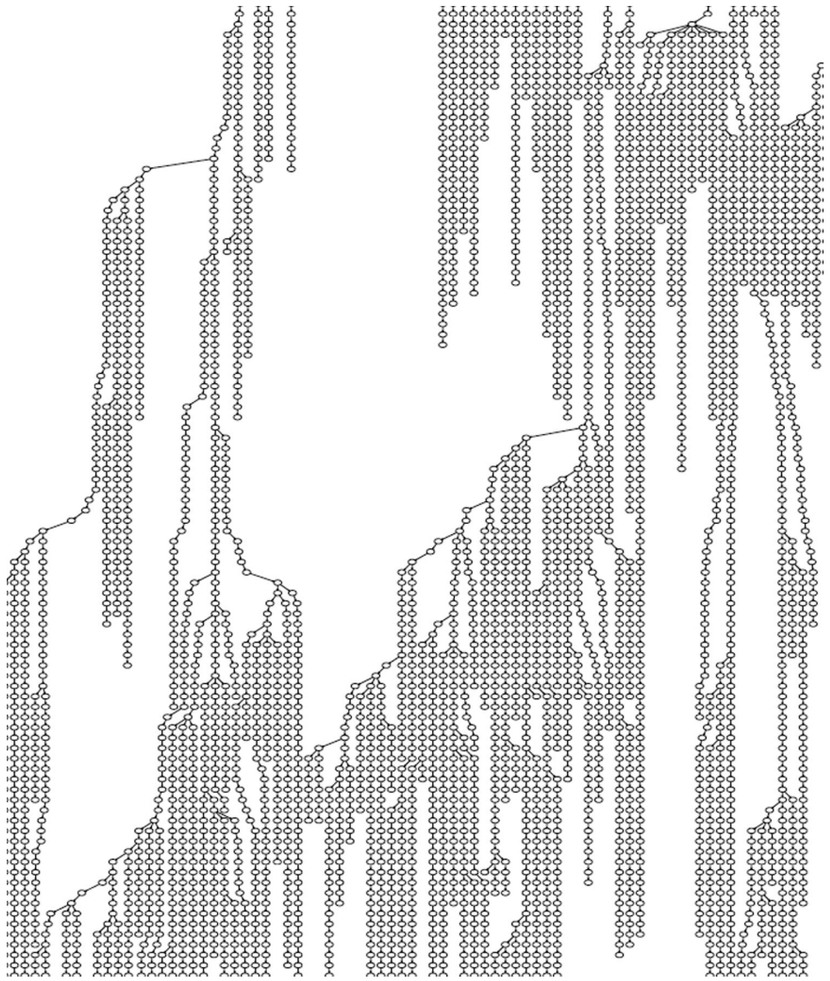
$p(\kappa)$

| $k$ | $\hat{p}(k)$ |
|---|---|
| 0 | 0.0246 |
| 1 | 0.9525 |
| 2 | 0.0217 |
| 3 | 0.0012 |
| ≥4 | 0 |

each node randomly draw a number of children from a given distribution

Is there a model that could fit the structures of the spreading process?

## Galton-Watson Branching Process



$p(\kappa)$

| $k$ | $\hat{p}(k)$ |
|---|---|
| 0 | 0.0246 |
| 1 | 0.9525 |
| 2 | 0.0217 |
| 3 | 0.0012 |
| ≥4 | 0 |

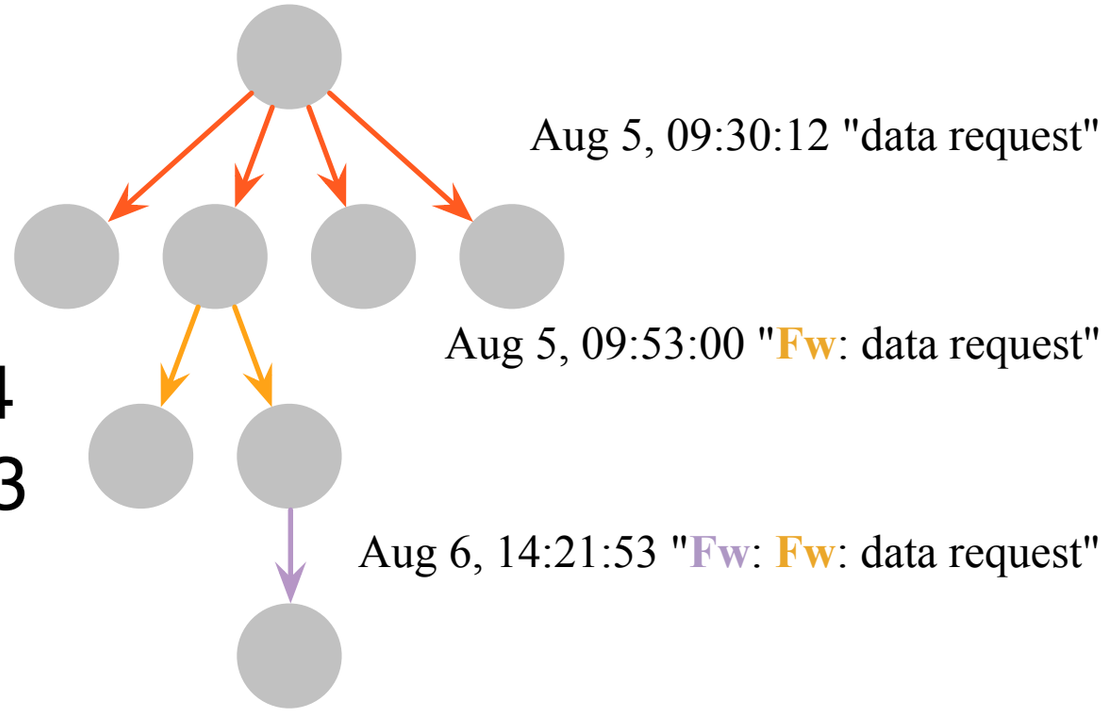
each node randomly draw a number of children from a given distribution

D. Liben-Nowell and J. Kleinberg, *Proc Natl Acad Sci* 2008

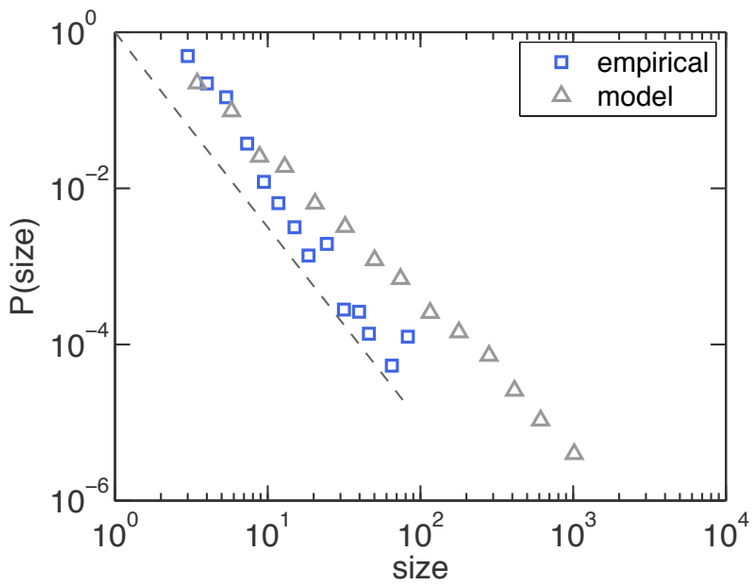
B. Golub and M. O. Jackson, *Proc Natl Acad Sci* 2010

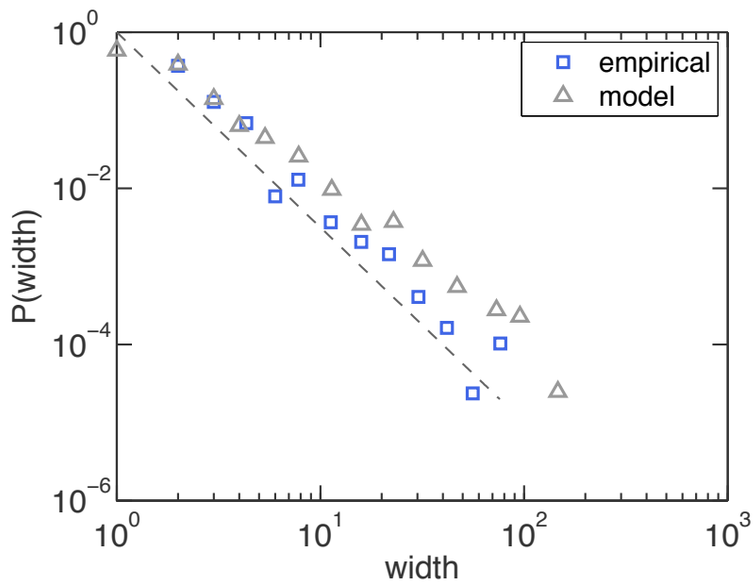
# Tree size, width, and depth

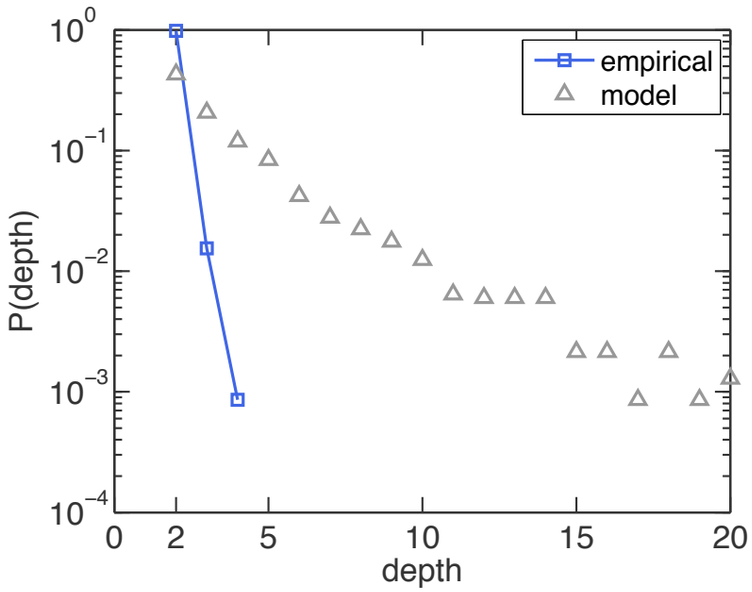
size = 8
width = 4
depth = 3



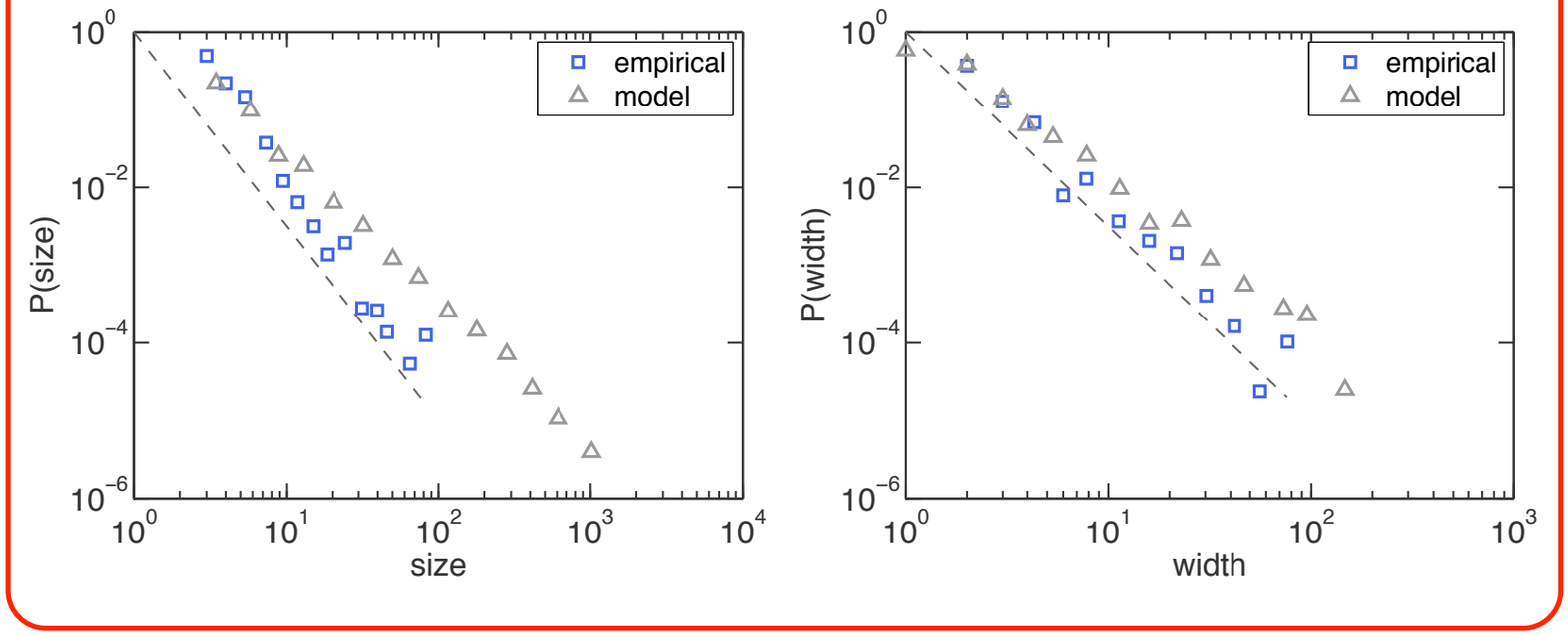

## Galton-Watson Branching Process



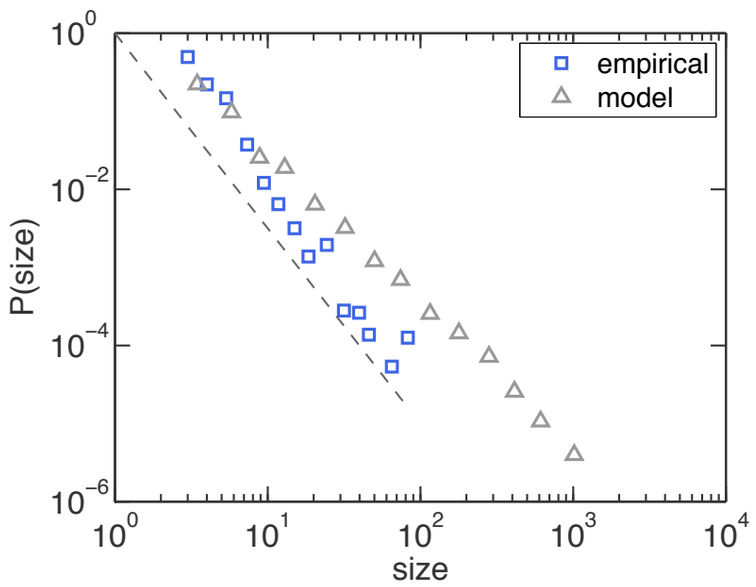



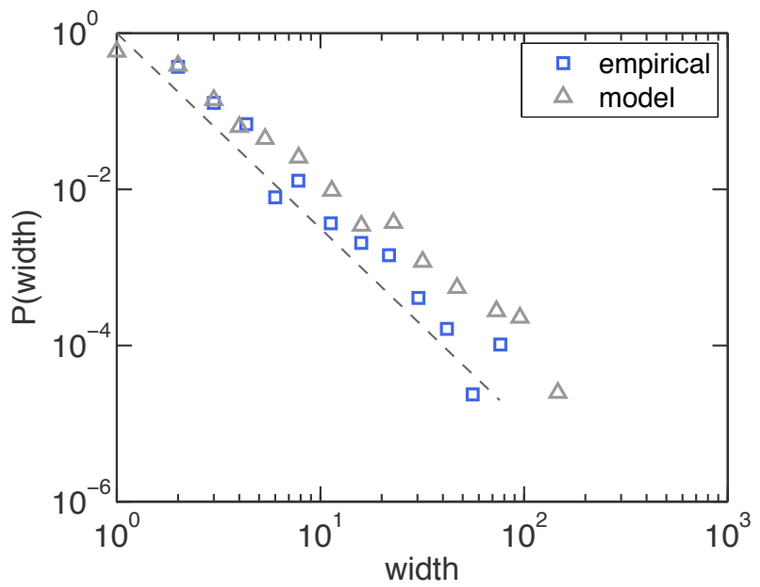



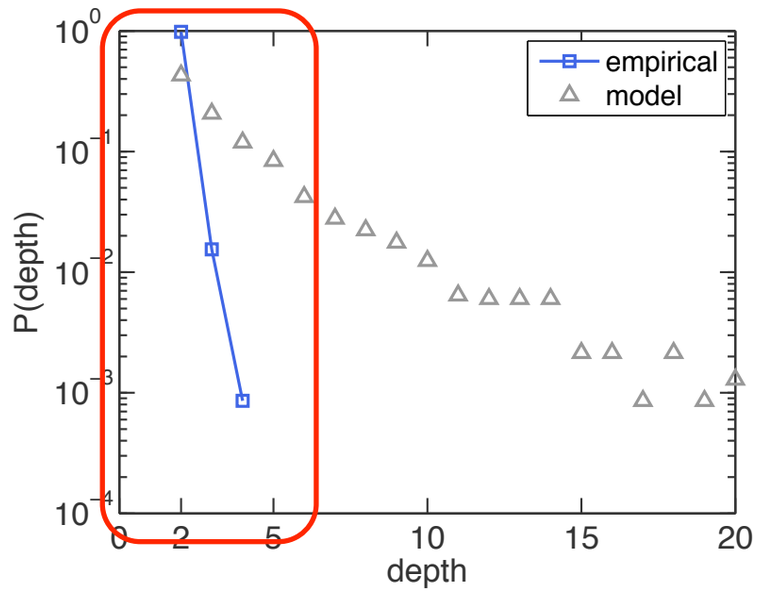
# Tree size, width, and depth



# Tree size, width, and depth

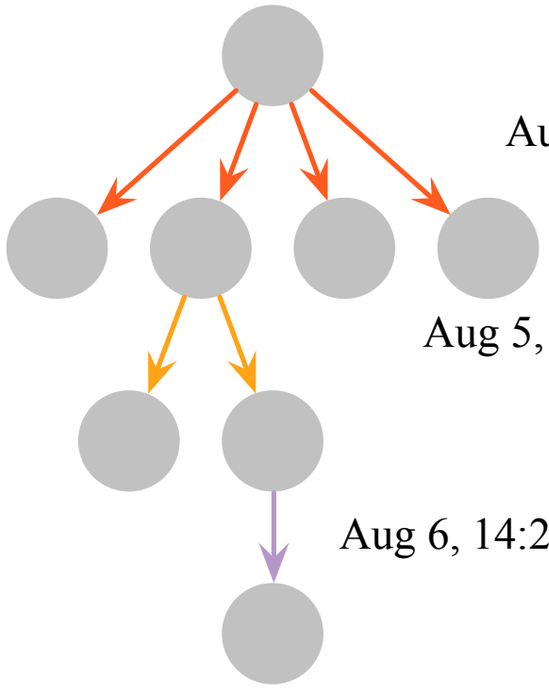



# Anomaly #1

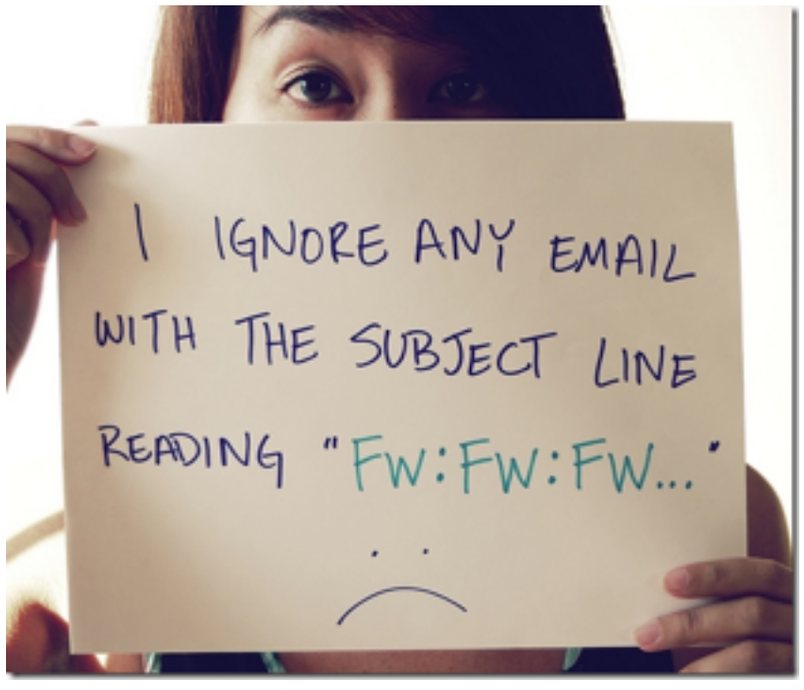
*Ultra Shallow!*



Aug 5, 09:30:12 "data request"

Aug 5, 09:53:00 "**Fw**: data request"

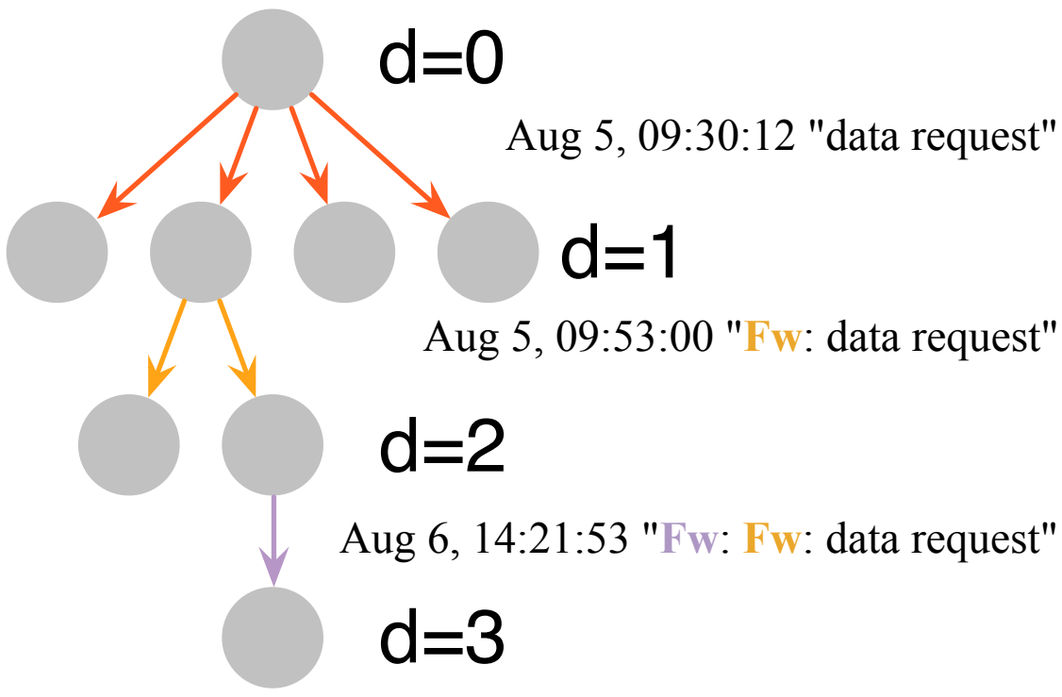
Aug 6, 14:21:53 "**Fw**: **Fw**: data request"



# Stage Dependence

$$P(\kappa \mid d)$$

## Galton-Watson Branching Process

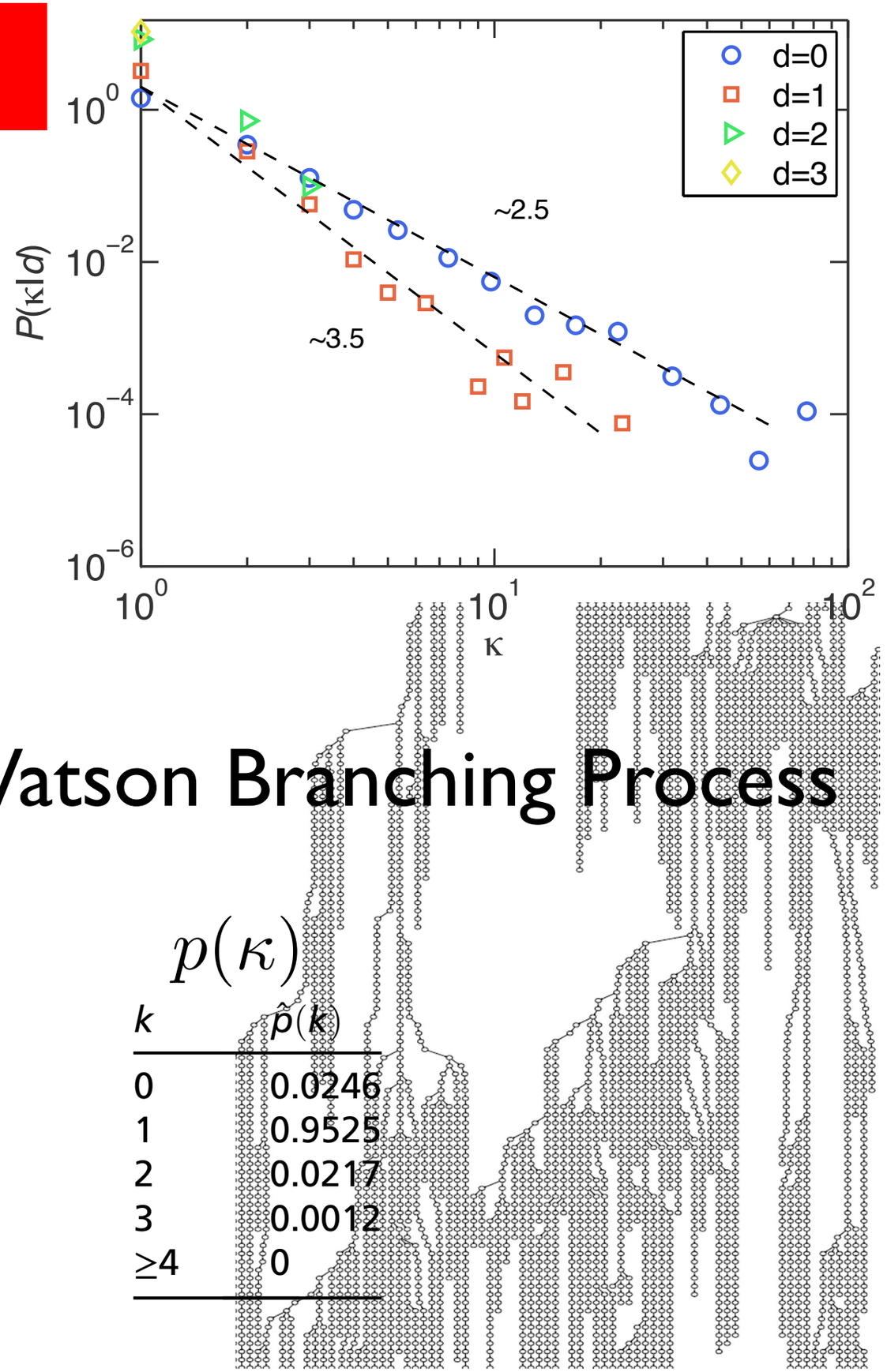



$p(\kappa)$

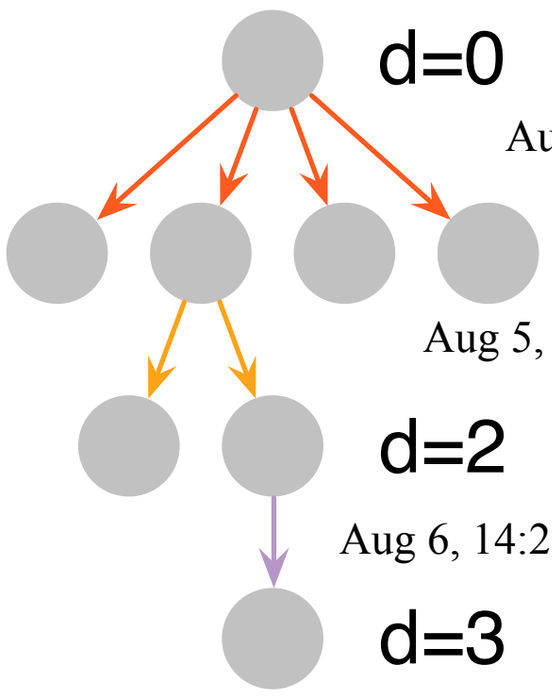
| $k$ | $\hat{p}(k)$ |
|---|---|
| 0 | 0.0246 |
| 1 | 0.9525 |
| 2 | 0.0217 |
| 3 | 0.0012 |
| ≥4 | 0 |

# Stage Dependence

$$P(\kappa \mid d)$$




# Galton-Watson Branching Process

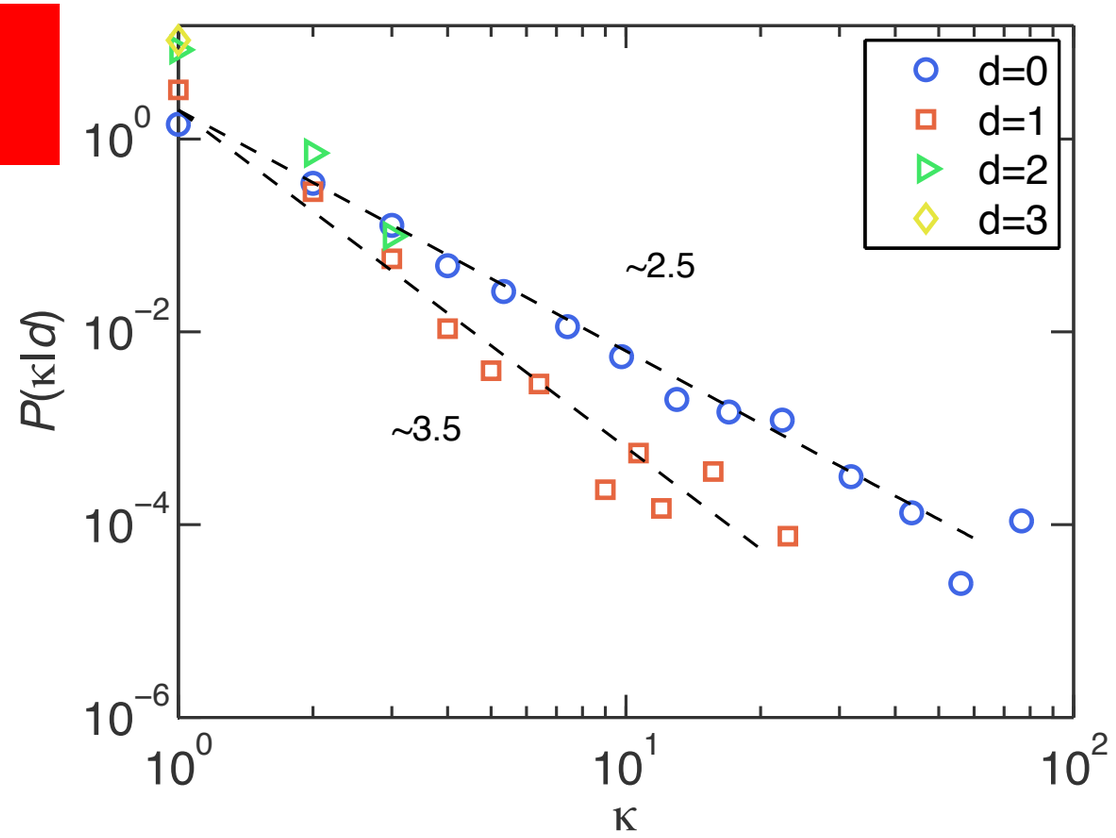



$$p(\kappa)$$

| $k$ | $\hat{p}(k)$ |
|---|---|
| 0 | 0.0246 |
| 1 | 0.9525 |
| 2 | 0.0217 |
| 3 | 0.0012 |
| ≥4 | 0 |

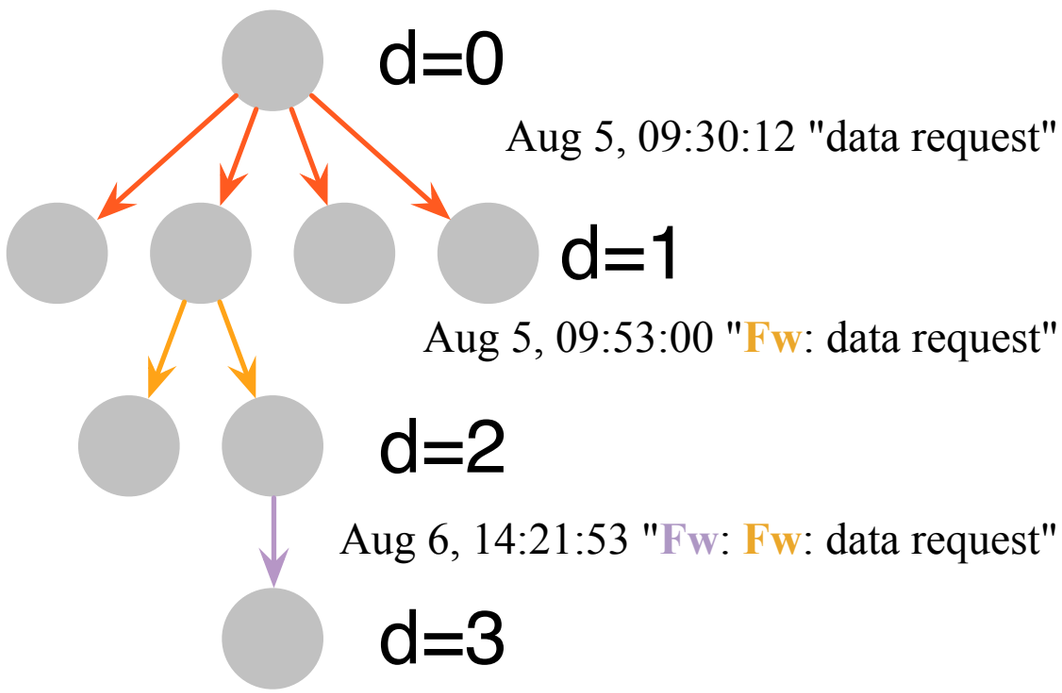
# Stage Dependence

$$P(\kappa \mid d)$$

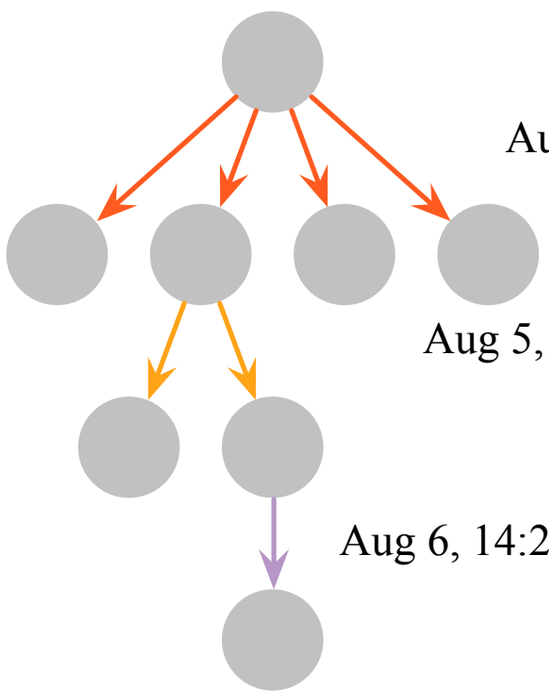



# Anomaly #2

## Stage Dependence




# Empirical Observations



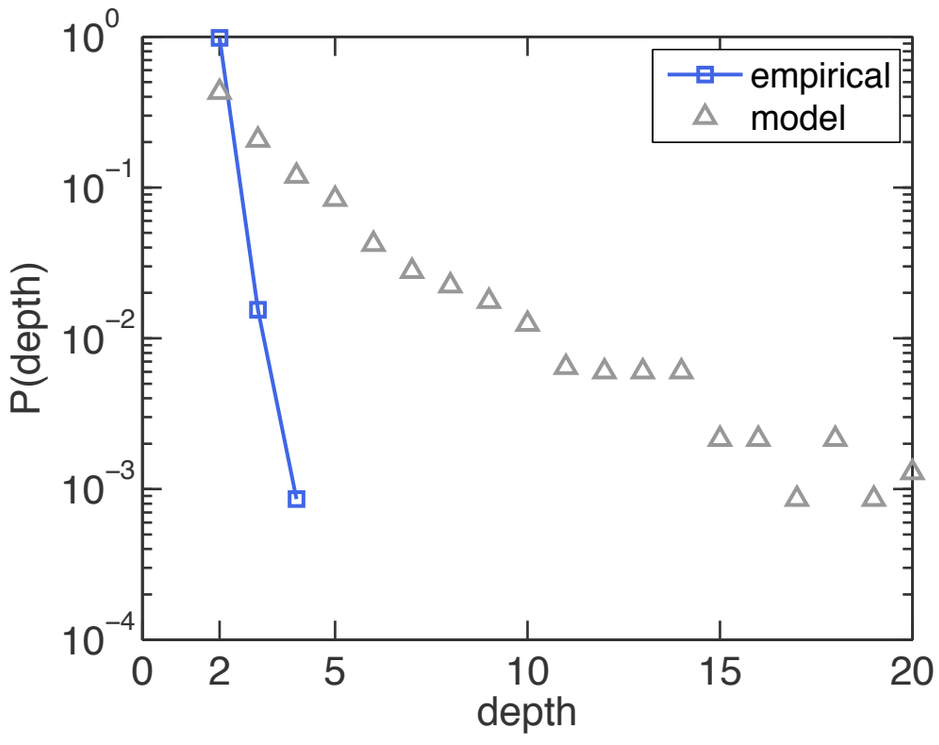
Aug 5, 09:30:12 "data request"

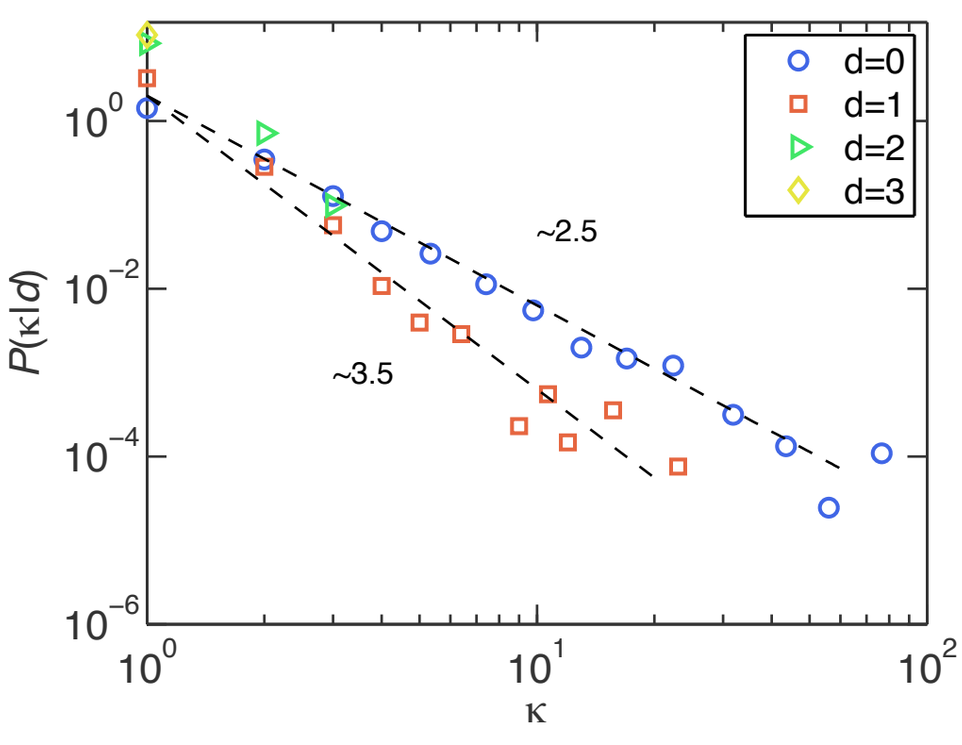
Aug 5, 09:53:00 "**Fw**: data request"

Aug 6, 14:21:53 "**Fw**: **Fw**: data request"

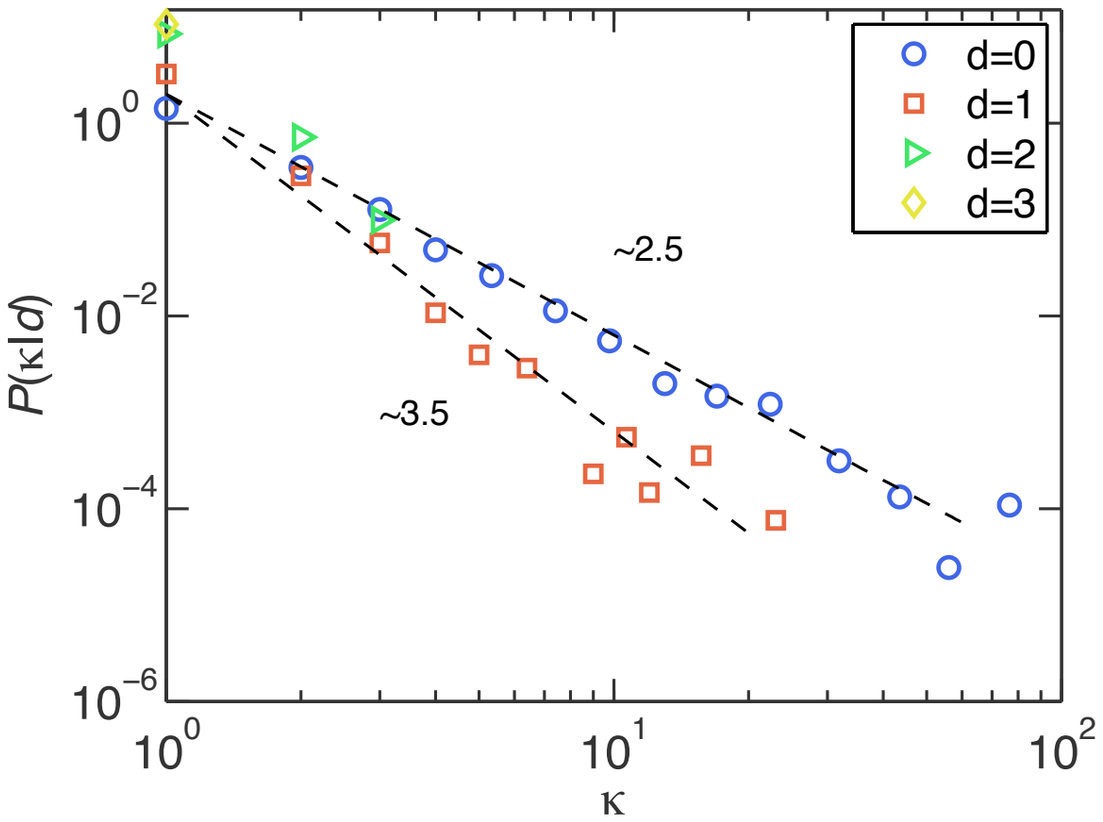
Ultra Shallow

Stage Dependence

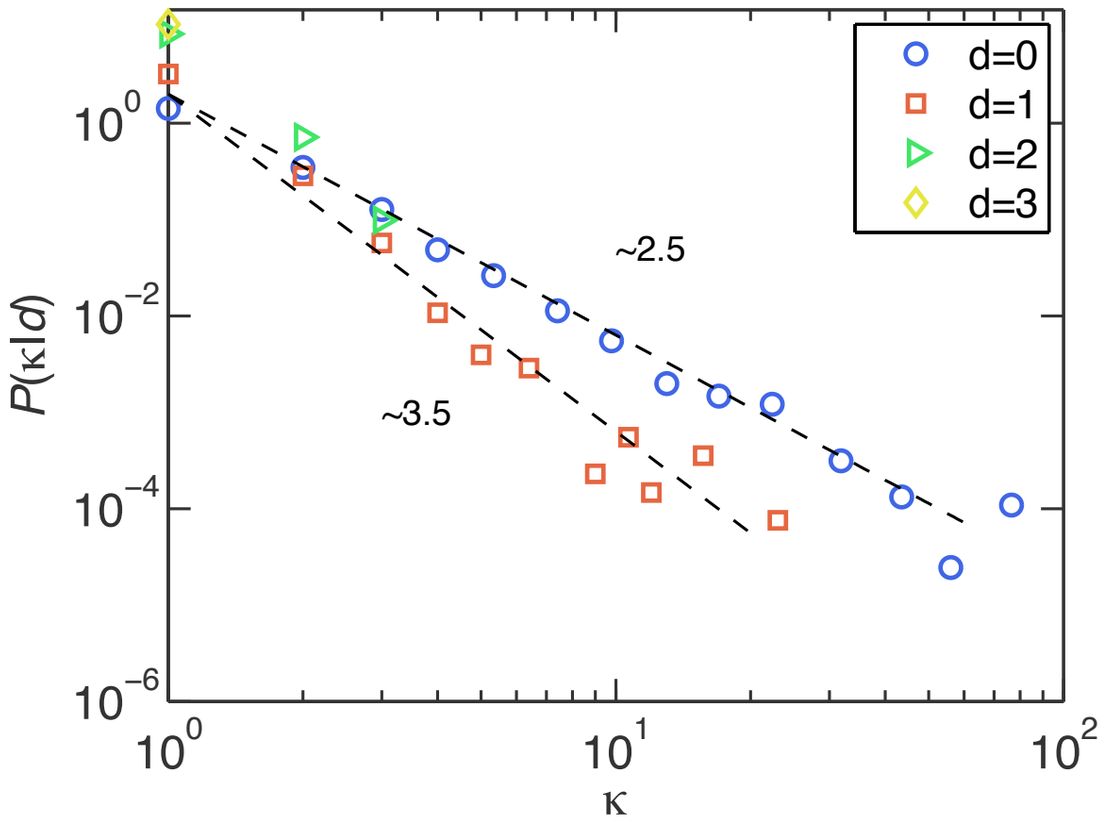


# Modeling the spreading processes



# Modeling the spreading processes

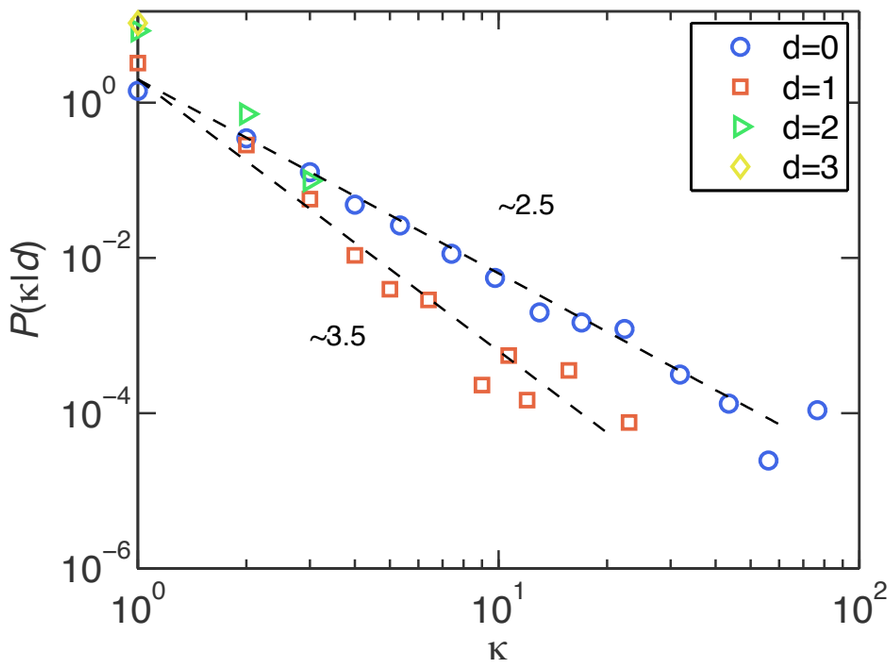



$$\kappa \; vs. \; k$$

$$P(\kappa) = \int P(\kappa \mid k) P(k) dk$$

$$P(\kappa) = P(\kappa \mid k) \quad \text{if independent}$$

# Modeling the spreading processes



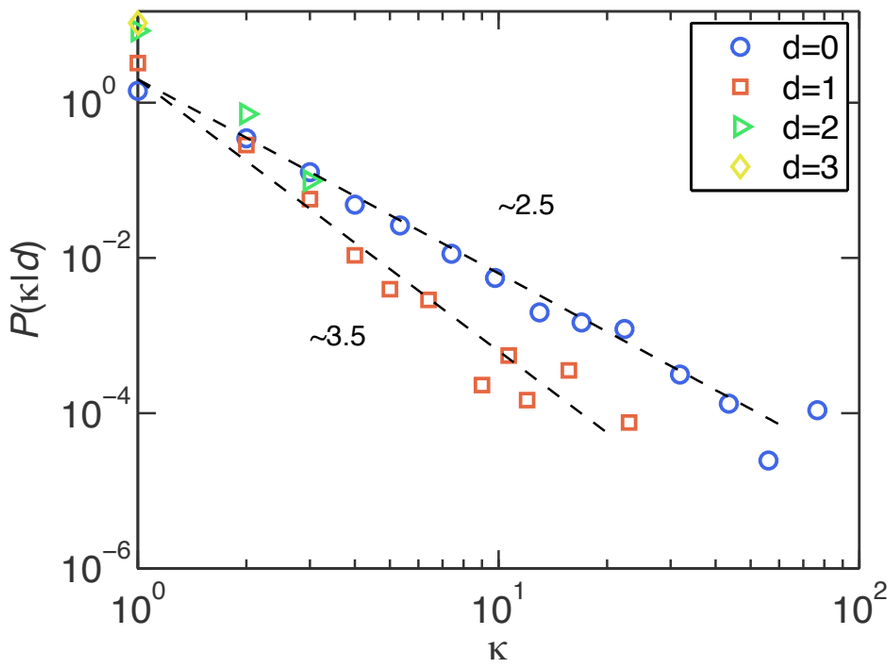

$$\kappa \ vs. \ k$$

$$P(\kappa) = \int P(\kappa \mid k) P(k) dk$$

$$P(\kappa) = P(\kappa \mid k) \quad \text{if independent}$$

- Modeling: disease vs. information
- Practice: choosing the seeds

# Modeling the spreading processes

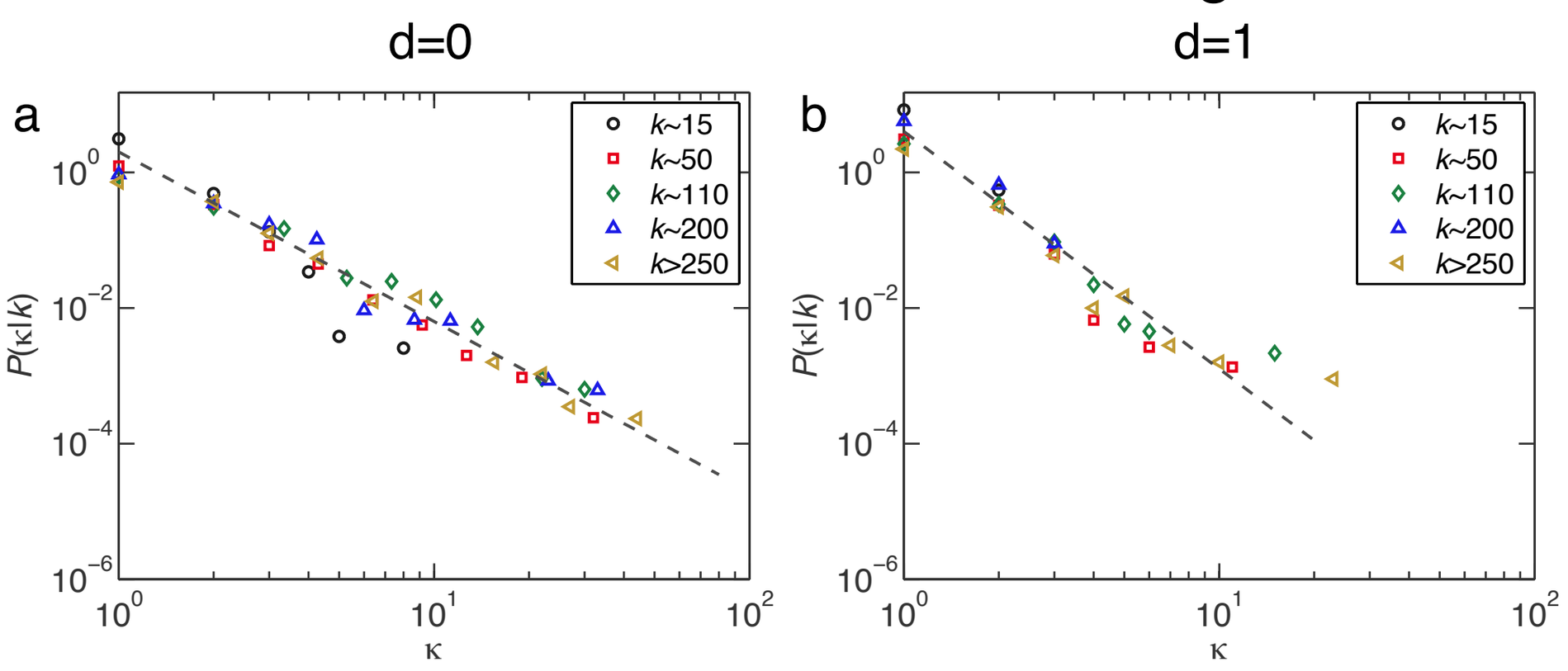



$$\kappa \ vs. \ k$$

$$P(\kappa) = \int P(\kappa \mid k) P(k) dk$$

$$P(\kappa) = P(\kappa \mid k) \quad \text{if independent}$$

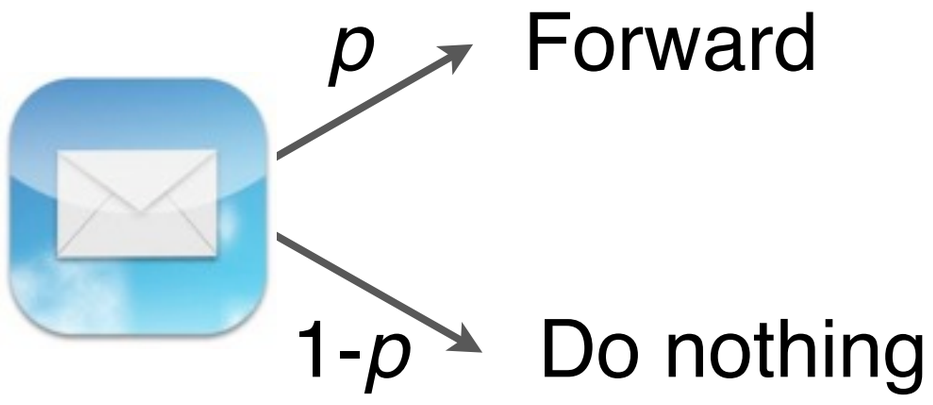
- Modeling: disease vs. information
- Practice: choosing the seeds



# Modeling the spreading processes

#recipients

$$P_n$$

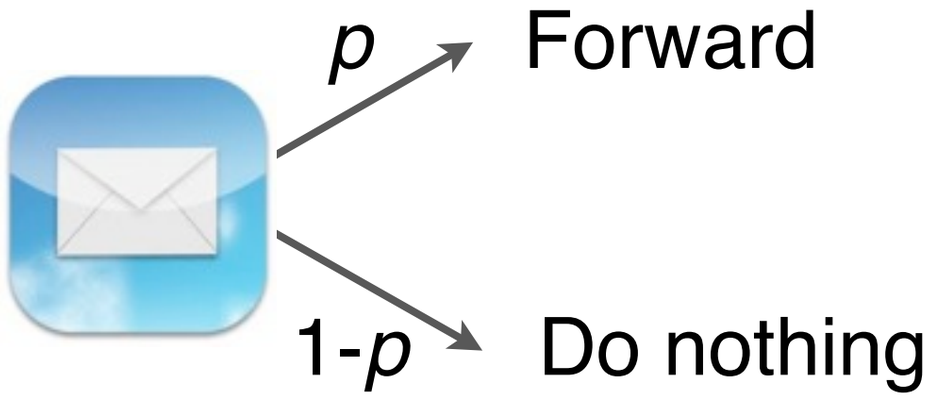



$$P(\kappa \mid d > 0) = P_n(\kappa)$$

# Modeling the spreading processes

#recipients

$P_n$




$$P(\kappa \mid d > 0) = P_n(\kappa)$$

when $d$=0

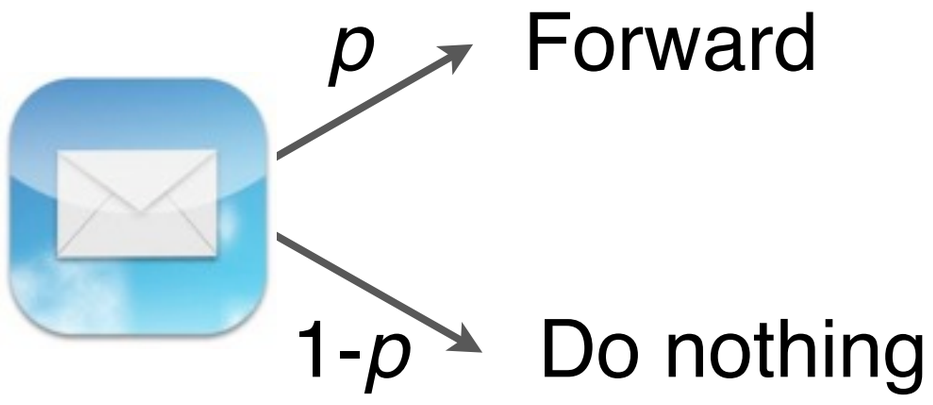
$$P(\kappa \mid d = 0) = A\left(1 - (1-p)^{\kappa}\right) P_n(\kappa)$$

$$= A\left(1 - e^{\kappa \ln(1-p)}\right) P_n(\kappa)$$

# Modeling the spreading processes

#recipients

$P_n$




$$P(\kappa \mid d > 0) = P_n(\kappa)$$

when *d*=0

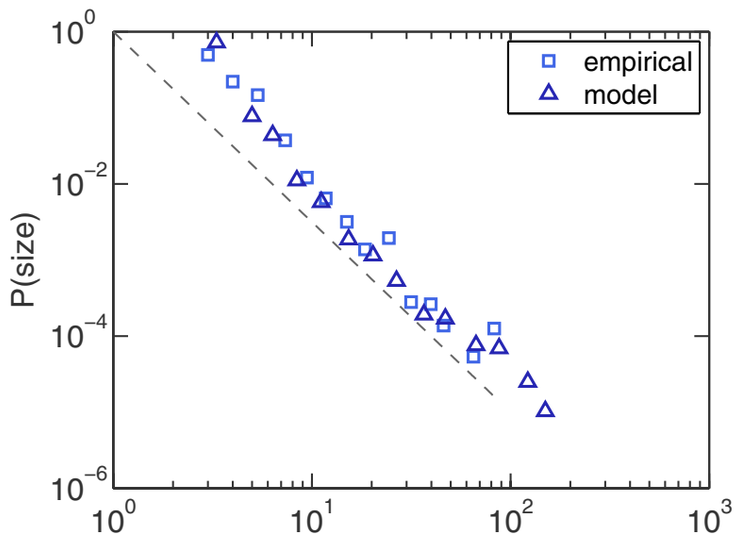
$$P(\kappa \mid d = 0) = A\left(1 - (1-p)^{\kappa}\right) P_n(\kappa)$$

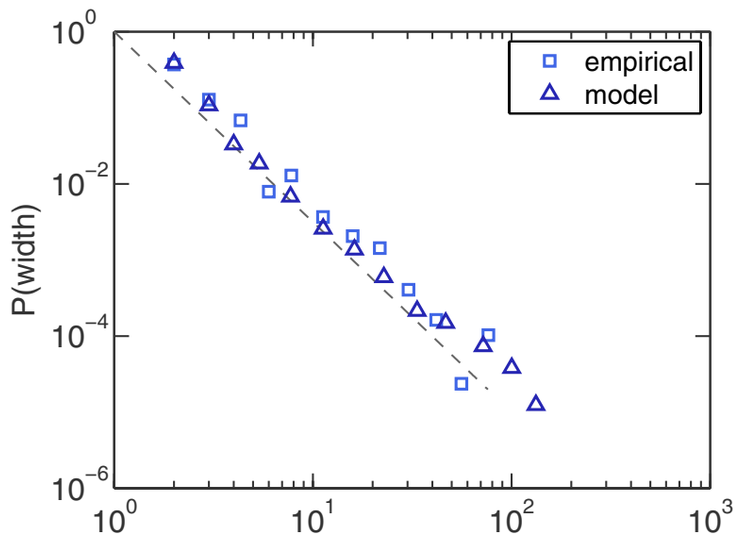
$$= A\left(1 - e^{\kappa \ln(1-p)}\right) P_n(\kappa)$$

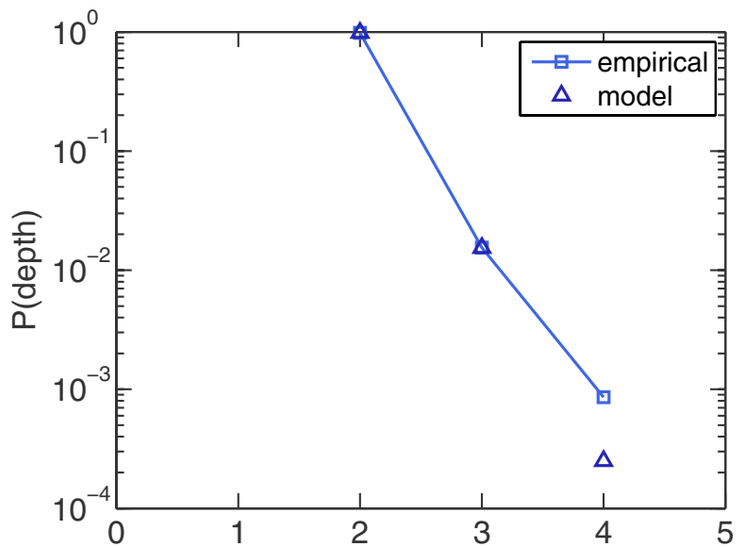
measured independently from the data

# Modeling the spreading processes

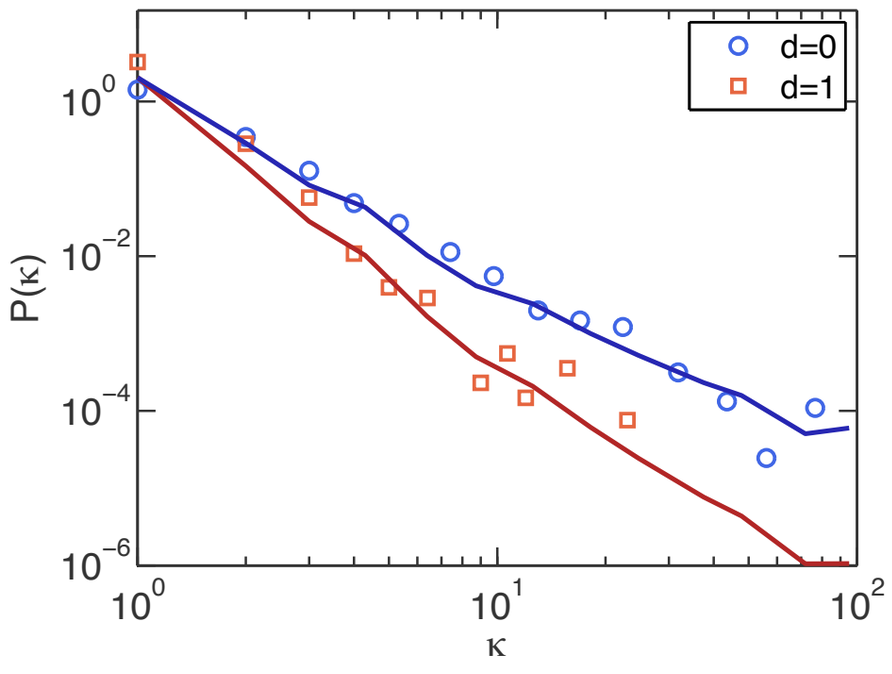
## Ultra Shallow




## Stage Dependence




a simple stochastic model captures a great deals of empirical observations

# Conclusion

- At the macroscopic level, the structures of spreading processes are largely independent of context.
- At the microscopic level, information spreading is indeed highly dependent on social context as well as individuals' behavioral profiles.

# Acknowledgement


Zhen Wen
Hanghang Tong
Ching-Yung Lin


Chaoming Song
Albert-László Barabási

Northeastern
HARVARD Medical School

**Information Spreading in Context**.
In *Proceedings of the 20th international conference on World Wide Web* (**WWW '11**)