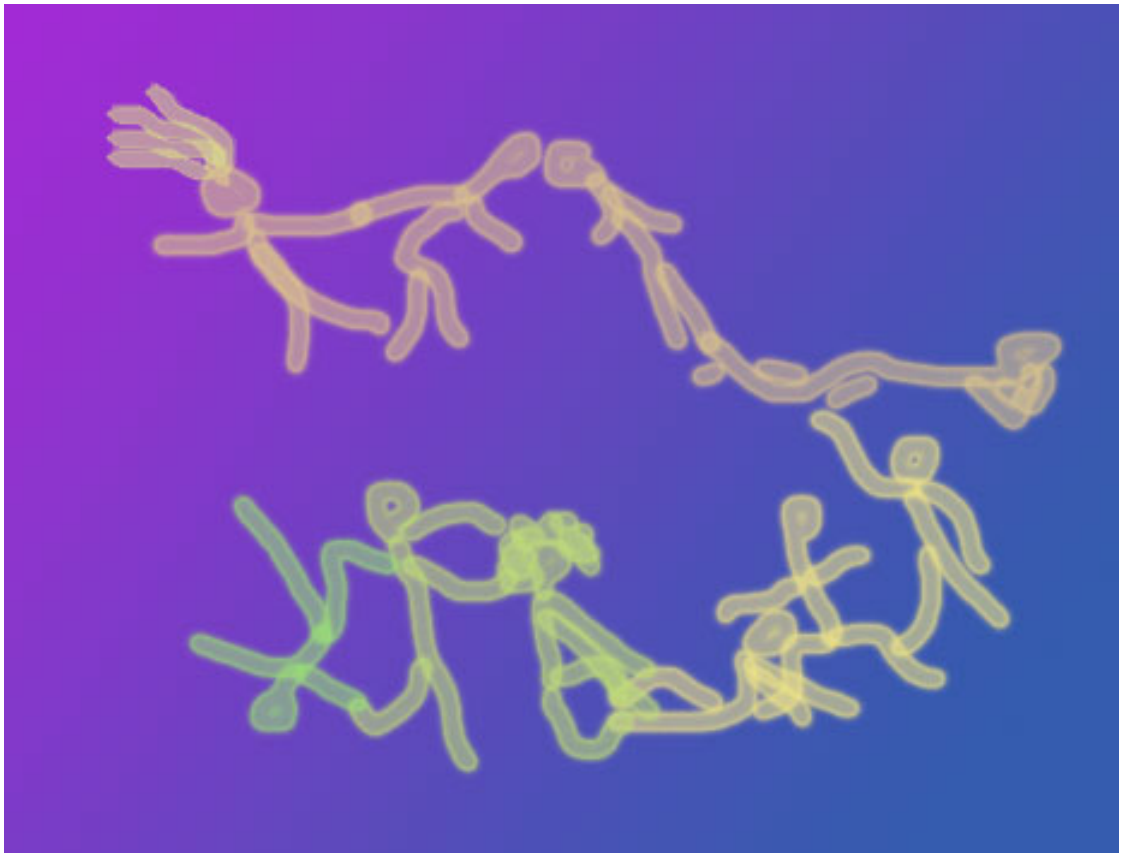


# Information Flow in Social Groups

Fang Wu, Bernardo A. Huberman, Lada A. Adamic, Joshua R. Tyler
HP Labs
Palo Alto, CA

# Outline

model and analytical results

email study

simulation on an email graph

# Information flow on a network

Information flow, like epidemics, is affected by the underlying network structure

Previous results on epidemics on scale free graphs:

Pastor-Satorras & Vespignani (2001) - no epidemic threshold in P-L graphs for exponents < 3

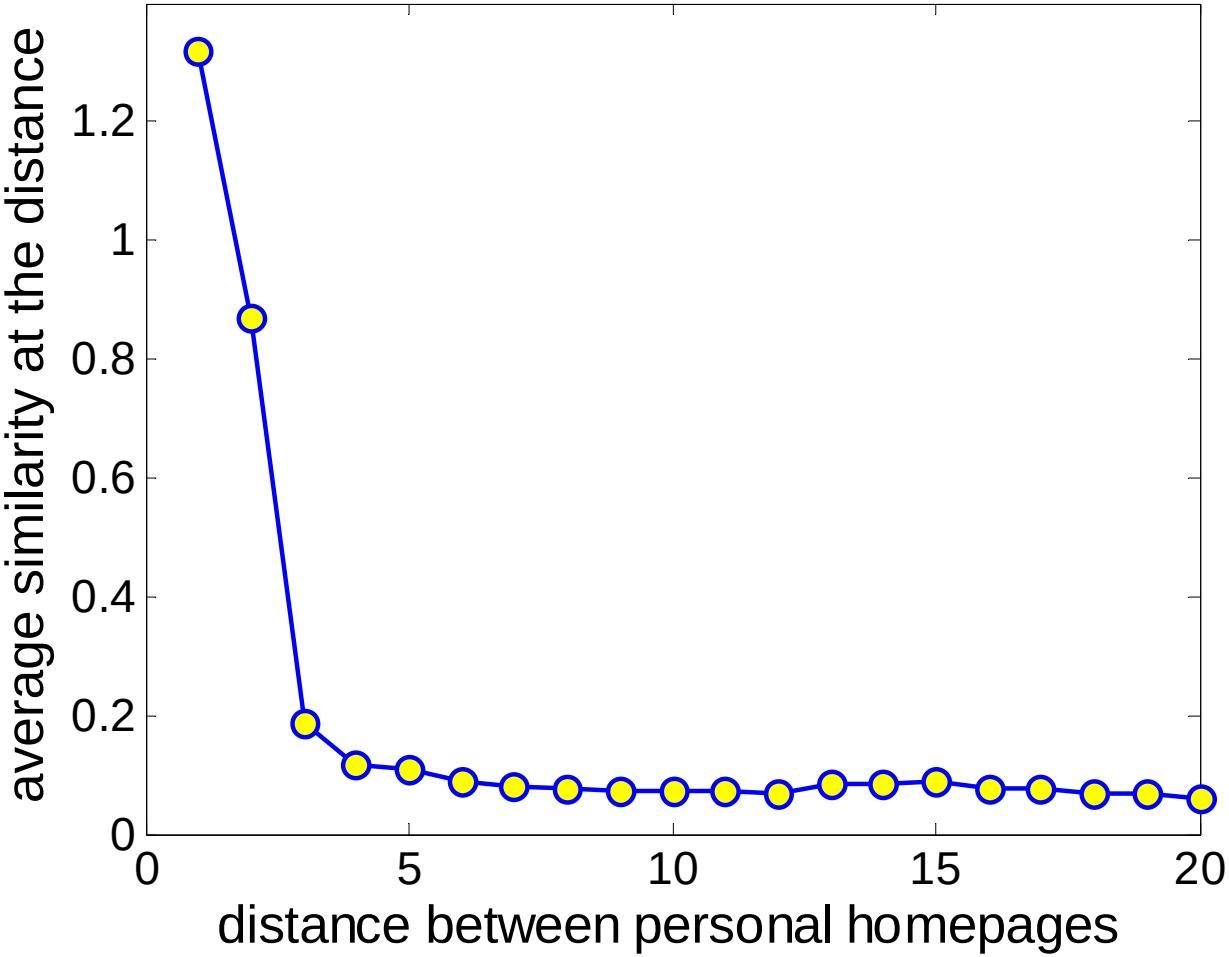
Newman (2002) - finite threshold in P-L graphs with an exponential cutoff

Newman (2002), Eguiluz & Klemm (2002), Vázquez & Moreno (2003) - finite threshold in P-L graphs with structure

## Additional factors affecting information spread

homophily: individuals with like interests associate with one another

textual similarity vs. # of hyperlinks separating homepages



# Modifying the basic SIR (Susceptible, Infected, Removed) model to reflect information as opposed to viral spread

Assumptions

1. Most information is not spread indiscriminately, but is passed selectively from one host to another based on knowledge of the recipient's interests

2. The further removed two individuals are in a network, the smaller the overlap in their interests (on average)

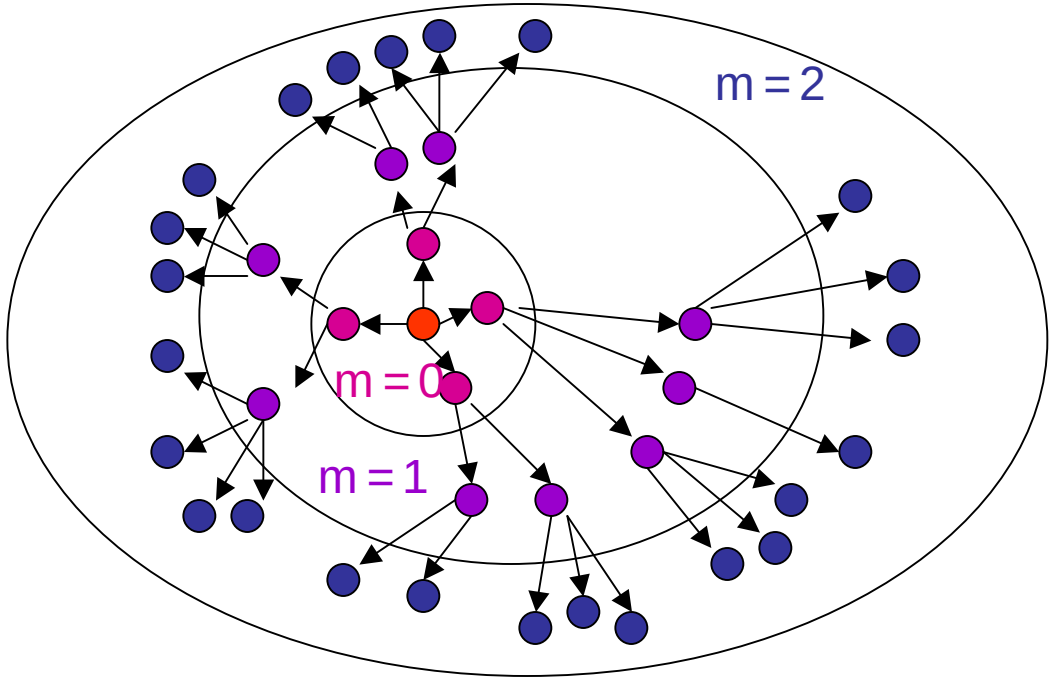
It follows that the likelihood that a piece of information will be transmitted decays as a function of the distance from the source node

# The Model

Decay in transmission probability as a function of the distance m between potential target and originating node

$$T^{(m)} = (m+1)^{-\beta} T$$

power-law implies slowest decay



# Analysis for a power-law network

Generating function calculation for the SIR model
(following Newman 2002)

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k \qquad p_k = k^{-\mathrm{a}} e^{-k/\mathrm{k}}$$

$$G_1(x) = \frac{G_0'(x)}{G_0'(1)}$$

condition that outbreak remains finite

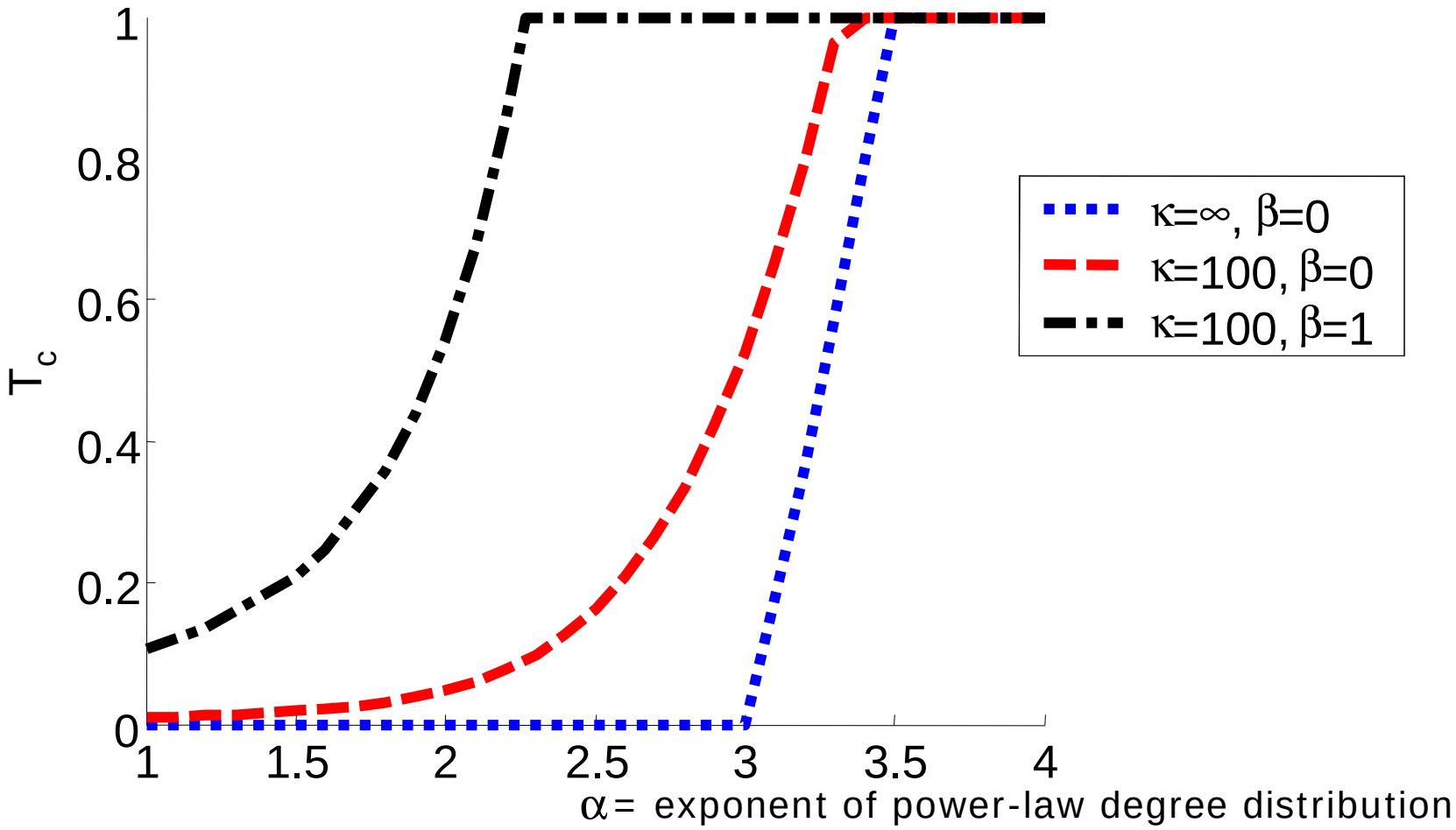
# of individuals receiving information at distance m+1 is
less than the number of individuals receiving the info at distance m

$$\frac{z_{m+1}}{z_m} = G_1^{(m)'}(1) = (m+1)^{-\mathrm{b}} T G_1'(1) < 1$$

always finite for some m, if $T^{(m)}$ decays with distance

# Effect of decay on $T_c$

$10^6$ nodes, epidemic if 1% ($10^4$) infected, $\beta = 1$, $\kappa = 100$
$T_c$ is the value of transmissibility above which epidemics occur



# Study of the spread of URLs and attachments

40 participants (30 within HPL, 10 elsewhere in HP & other orgs)

6370 URLs and 3401 attachments cryptographically hashed

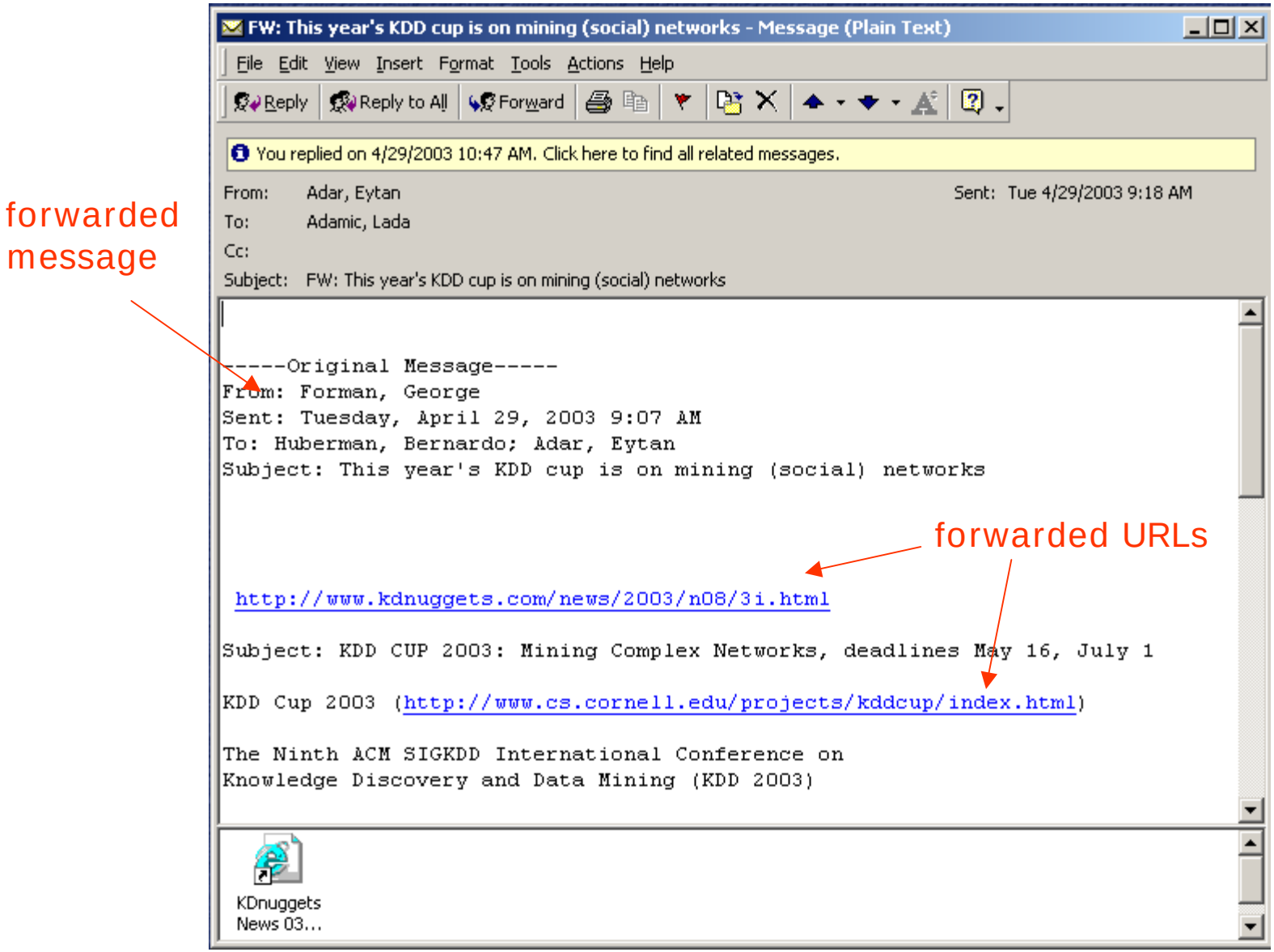
Question: How many recipients in our sample did each item reach?

caveats

messages are deleted (still, the median number of messages > 2000)

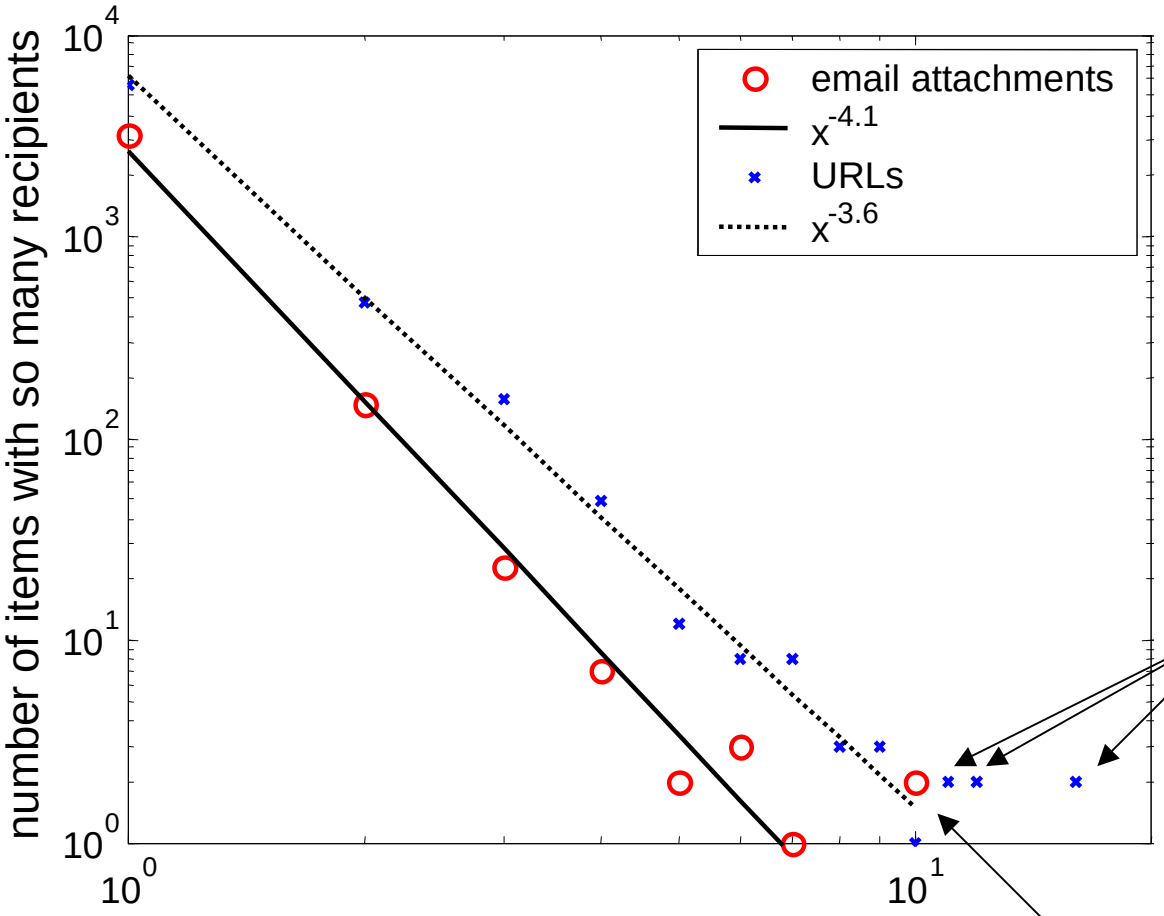
non-uniform sample (our research group is over-represented)

# Only forwarded messages are counted



# Results

average = 1.1 for attachments, and 1.2 for URLs




ads at the bottom of hotmail & yahoo messages

short term expense control

# Observations

Most URLs and attachments are not passed on more than once or twice

The ones that do reach a significant fraction of recipients are

- passed on involuntarily by the sender (advertisements embedded in email messages sent from free email accounts)
- passed top-down through the organizational hierarchy (an effective way to disseminate information that we do not account for here)

# Simulation of information transmission on the actual HP Labs email graph

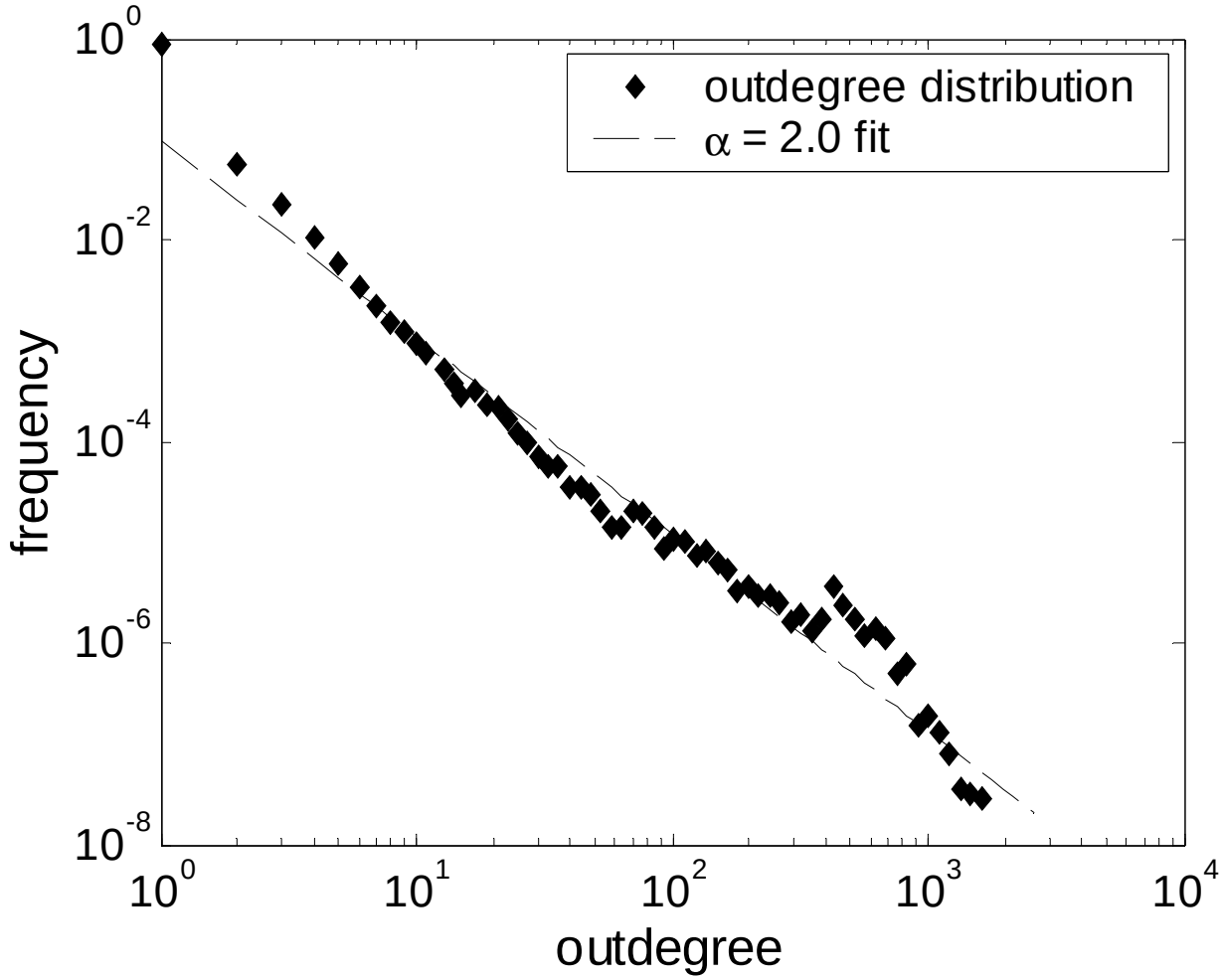
follow the HP Labs email log exactly: date, time, sender, and receiver (subject or content of email is not recorded)

120,000 entries involving > 7,000 potential recipients in 1 week

simulate how a piece of information would spread by doing the following:

- start by infecting one individual at random with a piece of information
- every time an infected individual sends an email they have a probability p of infecting the recipient
- individuals remain infected for 24 hours
- track epidemic over a week, most run their course in 1-2 days

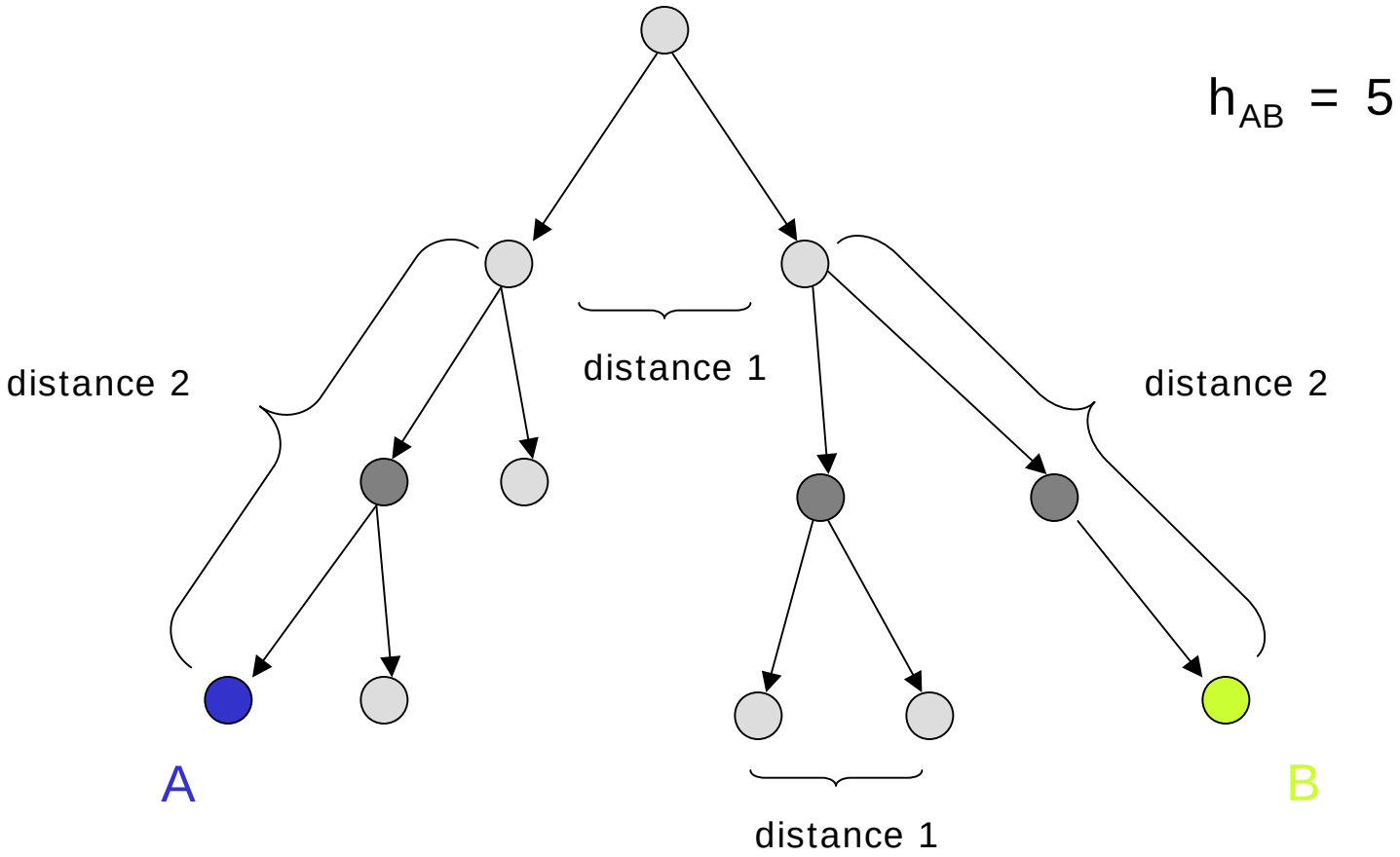
Degree distribution of all senders of email passing through the HP email server
(period = 3 months)




→ underlying network topology we are simulating is scale-free

Introduce a decay in the transmission probability based on the hierarchical distance

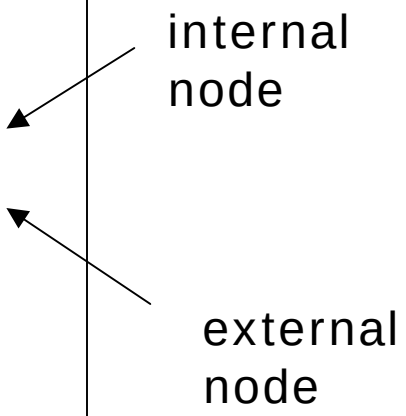
$$p = p_0 h^{-1.75}$$

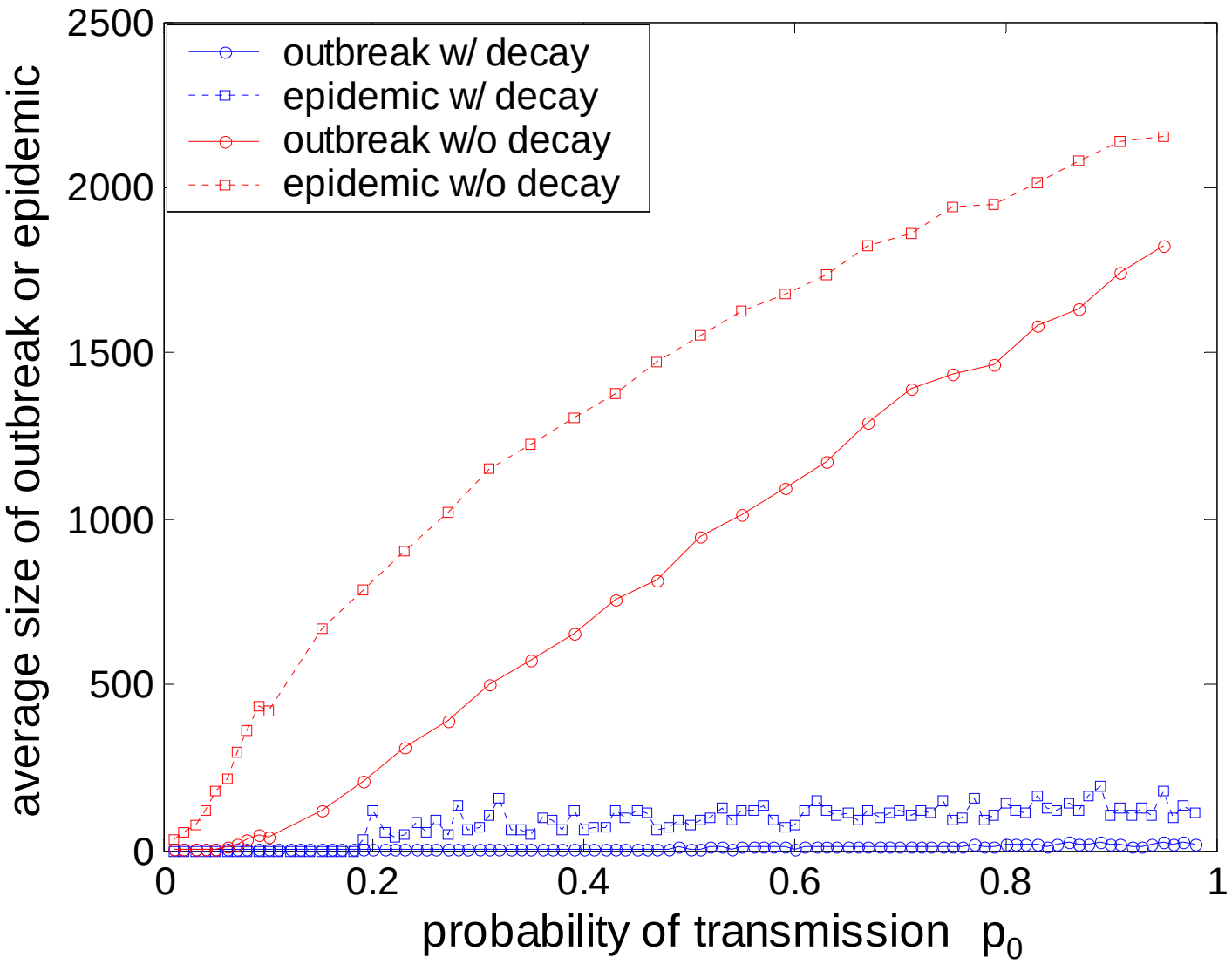


# Segment of email log used in simulation

each message has a probability p of transmitting information from an infected individual to the recipient

| | | | |
|---|---|---|---|
| 02/19/2003 | 15:45:33 | I-1 | I-2 |
| 02/19/2003 | 15:45:33 | I-1 | I-3 |
| 02/19/2003 | 15:45:40 | E-1 | I-4 |
| 02/19/2003 | 15:45:52 | I-5 | E-2 |
| 02/19/2003 | 15:45:55 | E-3 | I-6 |
| 02/19/2003 | 15:45:58 | I-7 | I-8 |
| 02/19/2003 | 15:46:00 | E-4 | I-9 |
| 02/19/2003 | 15:46:05 | I-10 | I-11 |
| 02/19/2003 | 15:46:10 | I-12 | I-13 |
| 02/19/2003 | 15:46:10 | I-12 | I-14 |
| 02/19/2003 | 15:46:10 | I-12 | I-15 |
| 02/19/2003 | 15:46:14 | I-16 | E-5 |
| : | : | : | : |




Simulated reach of information
outbreak w/ decay
epidemic w/ decay
outbreak w/o decay
epidemic w/o decay
average size of outbreak or epidemic
2500
2000
1500
1000
500
0
0
0.2
0.4
0.6
0.8
1
probability of transmission $p_0$

# Comments on simulations results

In the absence of transmissibility decay

- epidemics occur even for low transmissibility ($p_0 = 0.01$).
- on average, outbreaks affect a sizable fraction of the potential recipients

When decay is present

- No epidemics occur below $p_0 = 0.2$ (i.e. there is a threshold).
- Even for high baseline transmissibility, outbreaks are contained to a small fraction of potential recipients

# Sample simulations
## with p = 0.18

Animation of simulated information spread without decay
available as AVI (5.1M) and MPEG (1.3M)
information spreads through roughly half of the network

Animation of simulated information spread with decay
available as AVI (681K) and MPEG (96K)
information is contained to the group of the original source

Animations were created using Zoomgraph, an open source package for visualizing networks. It can be obtained from http://www.hpl.hp.com/shl/projects/graphs/

# Conclusions

Under the assumptions that

- Information is passed on to individuals with matching properties
- The likelihood that properties match decreases with distance from the source

Model gives a finite threshold
Information spread typically does not reach epidemic proportions

Results are consistent with observed URL & attachment frequencies in a sample

Simulations following real email patterns also consistent

preprint is available at http://www.hpl.hp.com/shl/papers/flow/