



# Implicit Structure and Dynamics of Blogspace

Eytan Adar  
May 18, 2004

(joint work with: Li Zhang, Lada Adamic, and Rajan Lukose)

© 2004 Hewlett-Packard Development Company, L.P.  
The information contained herein is subject to change without notice



# Blogs and the digital experience

- Significant use: online diary
  - Record real-world and virtual experiences
  - Easy to record things “seen” on the net
- Structure: blog-to-blog linking
- Use + Structure
  - Great to track “memes”

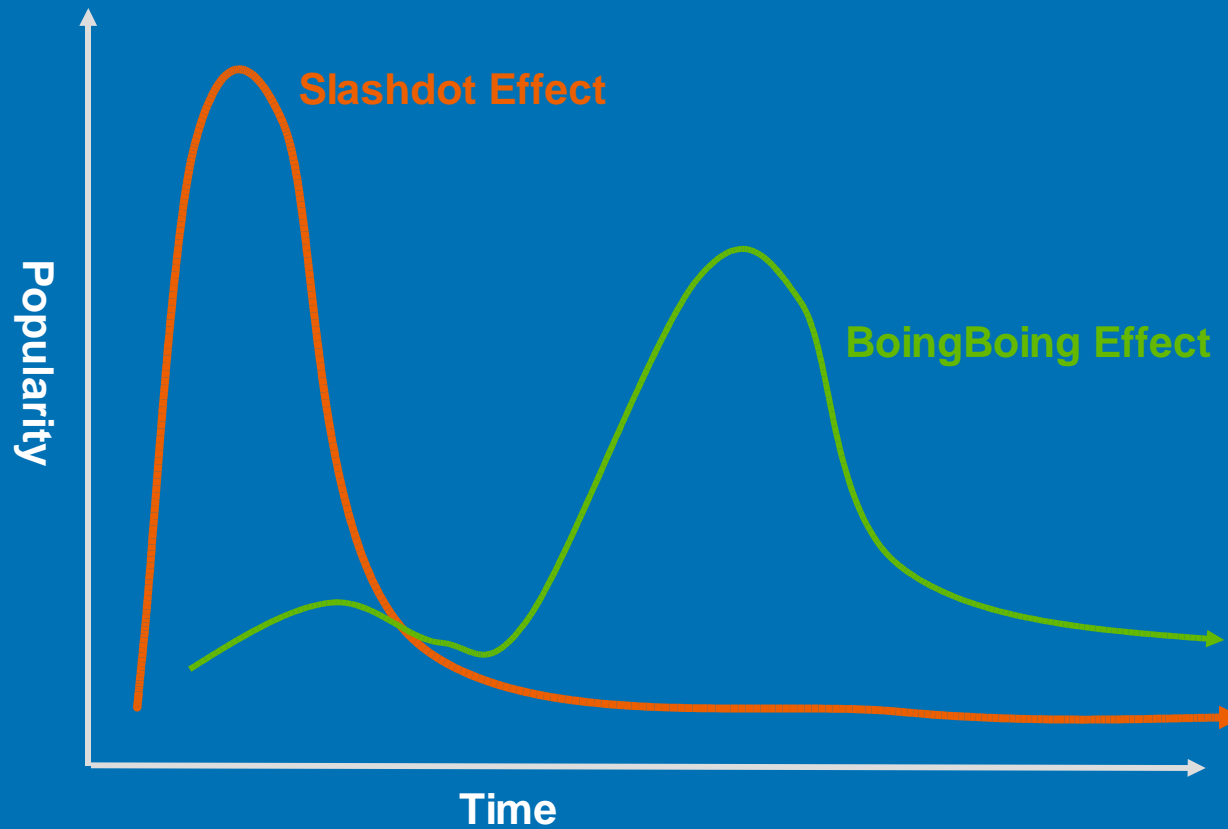
# Our interest

- Macroscopic patterns of blog epidemics
- Microscopic patterns of blog epidemics
  - Implicit & Explicit
- Ranking algorithms that take advantage of infection patterns

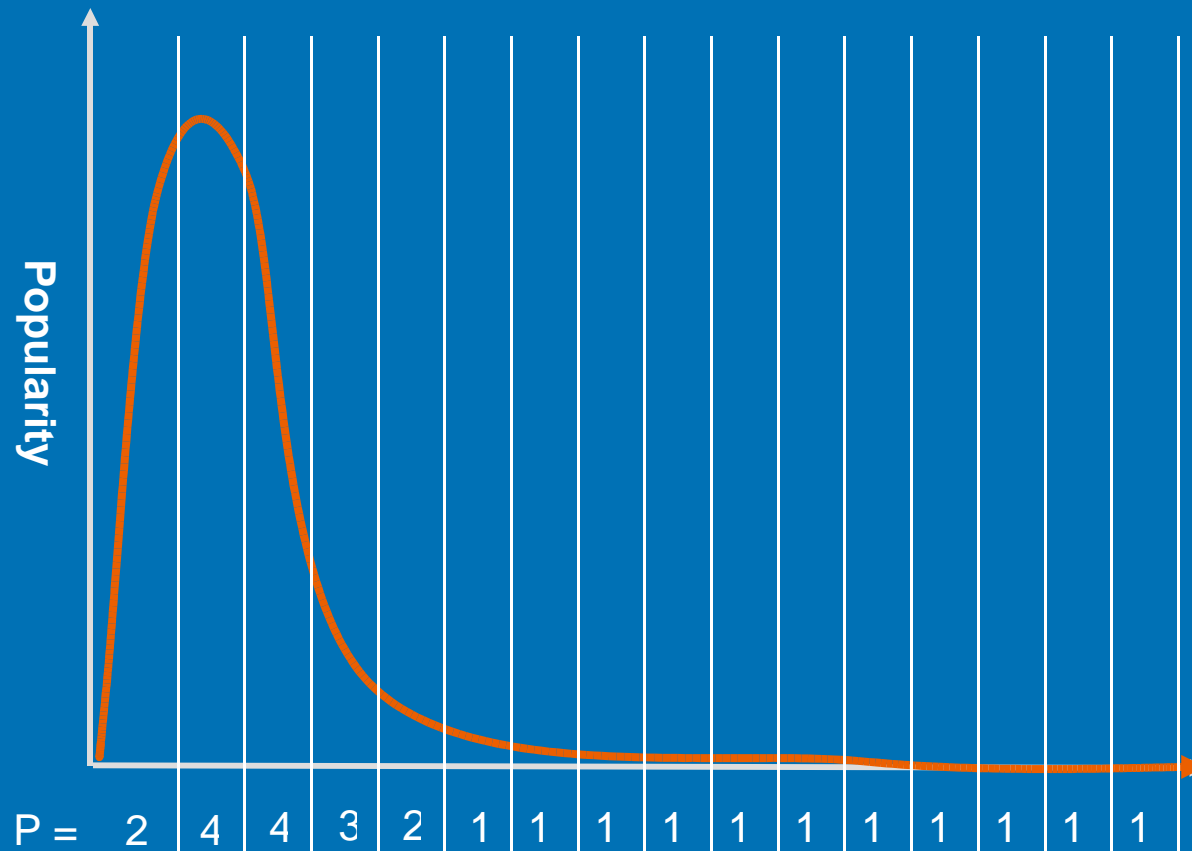
# Tracking Blogs

- Blogdex: Earliest example
  - Lets you see which blogs (and when) linked to a site
  - Others emerged with similar/related functionality
- Can find epidemic profiles (popularity over time)
- Our question: do different types of information have different epidemic profiles?
- Conversely: do different types of epidemic profiles group similar types of information?

# For Example...



# Clustering Epidemic Profiles



# Clustering Epidemic Profiles

- K-means clustering (on normalized data)
  - Assume that data clusters into k groups
    - (we tried a few ways, 4 worked best)
  - 259 URLs

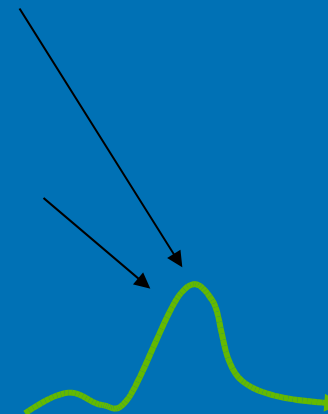
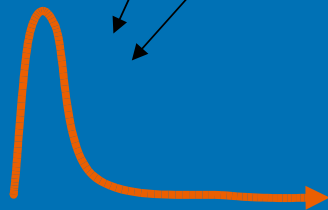
$$P_1 = [2\ 4\ 4\ 3\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1]$$

$$P_2 = [1\ 1\ 1\ 1\ 1\ 2\ 2\ 3\ 3\ 2\ 2\ 1\ 1\ 1\ 1\ 1]$$

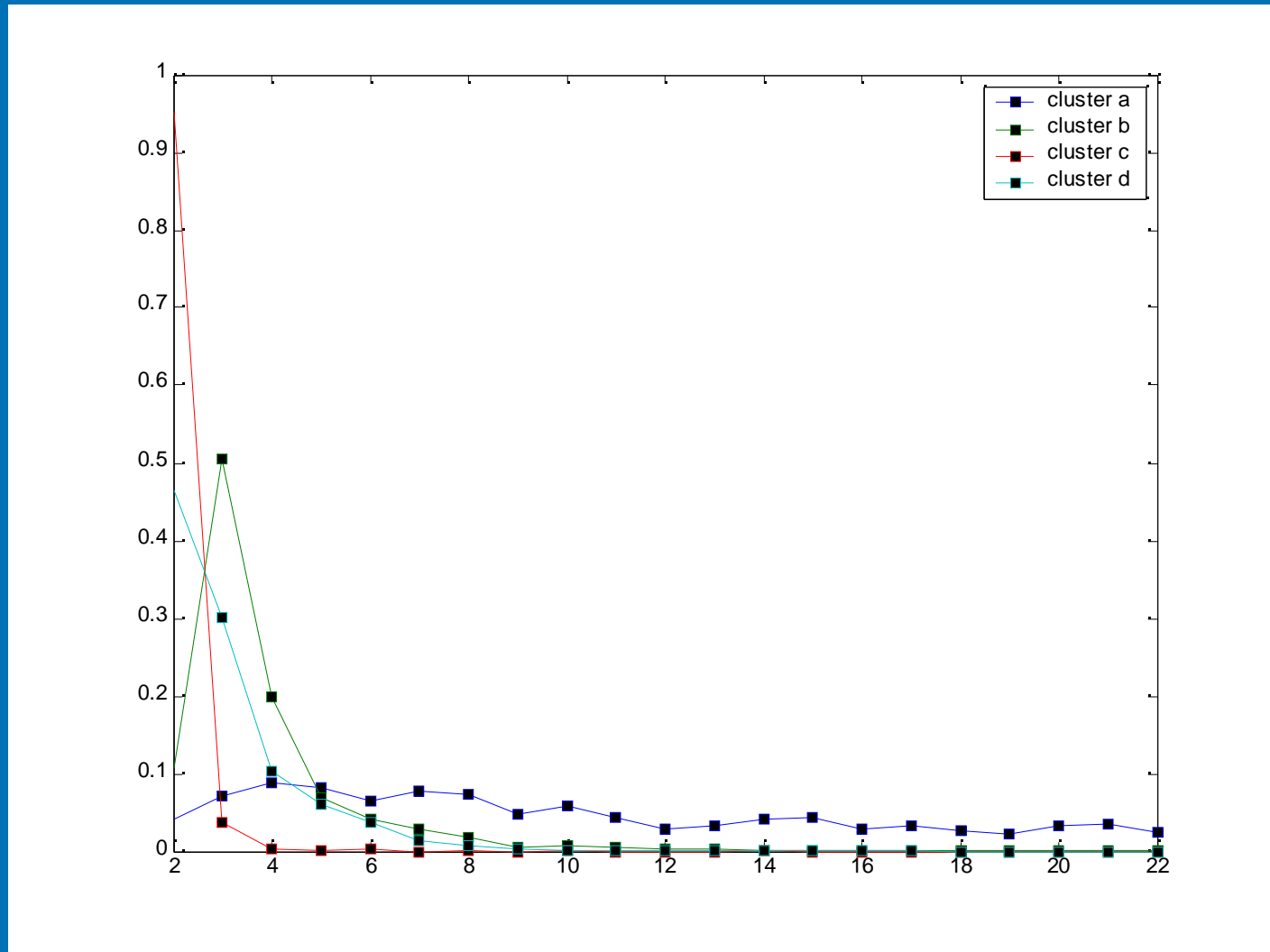
$$P_3 = [2\ 5\ 4\ 3\ 2\ 1\ 1\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 1\ 1]$$

$$P_4 = [1\ 1\ 1\ 1\ 1\ 3\ 3\ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 1]$$

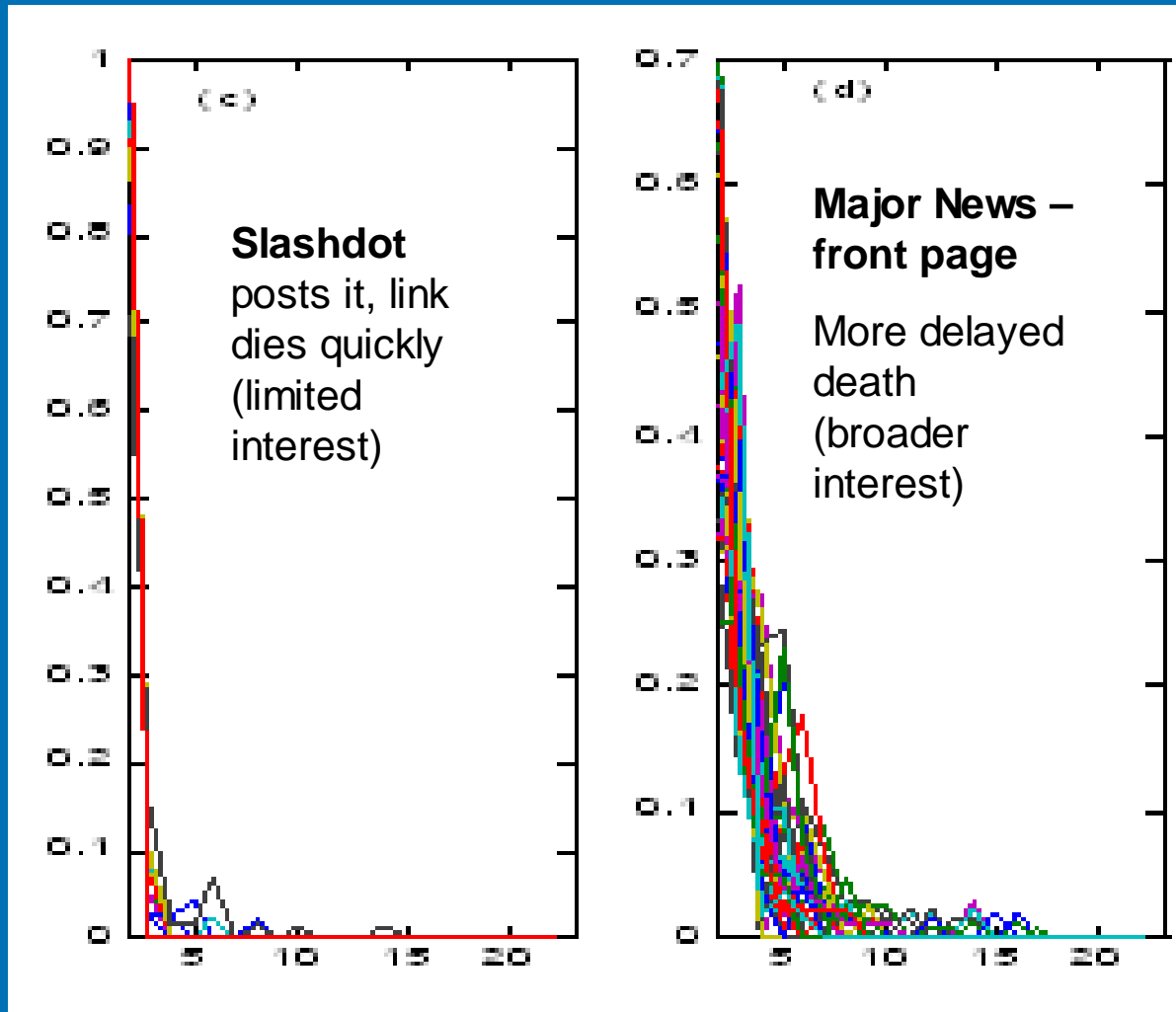
..



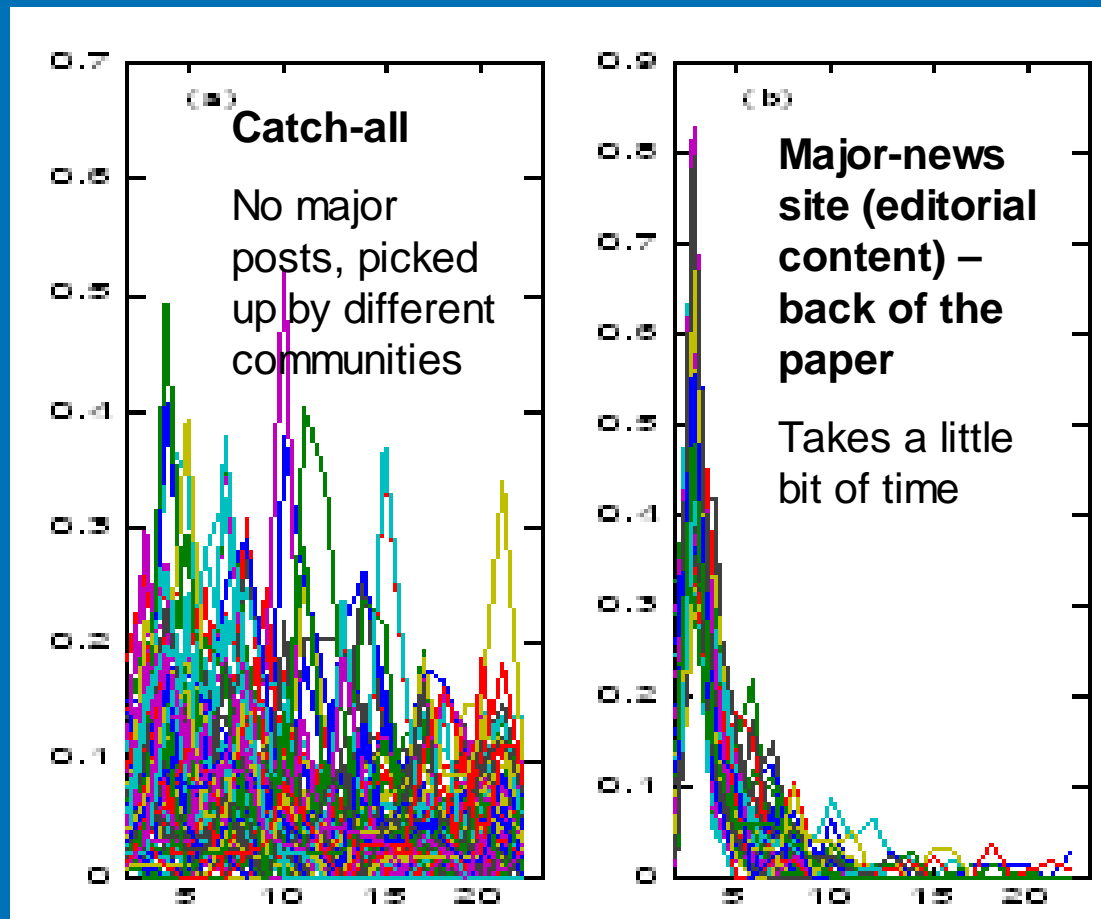
# Centroids



# Clusters



# Clusters



# Microscale Example: Giant Microbes

5.10.2003

Just what our children need: [common diseases made cute](#), posted at 11:25 AM by rachel

Embrace illness with [Giant Microbes plush toys](#)! Man, every med student from here to Kalamazoo is going to be inundated with these tchotchkes come Christmas-time. Via [Roninneko](#).  
posted by Reen | [link](#) | ...talkety...4 comments

[lar](#)) • Also cute, and [animals that look](#)

[like tiny microbes](#)—only a million times actual size! Now available: The Common Cold, The Flu, Sore Throat, and Stomach Ache." And coming soon: [Martian microbes](#) (via [Reenhead](#)) • [Lou Reed](#) on the cover of Kung

microbes—only a million times actual size! now available: The Common Cold, The Flu, Sore Throat, and Stomach Ache. "  
[link via Incoming Signals](#)



Posted by nchicha at May 14, 2003, 04:50 AM |

medpundit

Commentary on medical news by a practicing physician.

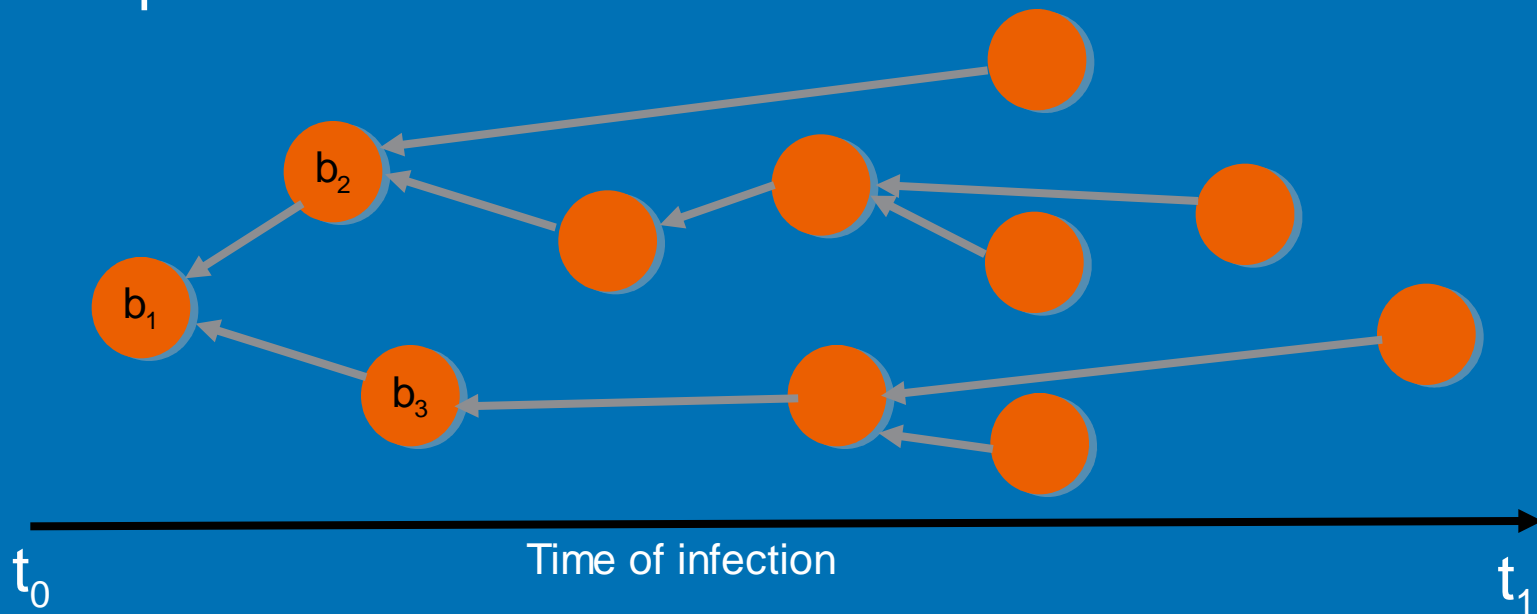
Friday, May 14, 2003

**Plush Bugs:** Looking for the perfect gift for that special someone? Try [GIANTmicrobes](#)!. They've got [influenza](#), the [common cold](#), [shigella](#) (which causes diarrhea), and [strep throat](#). (via [Cup of Chicha](#))

pico search  Search

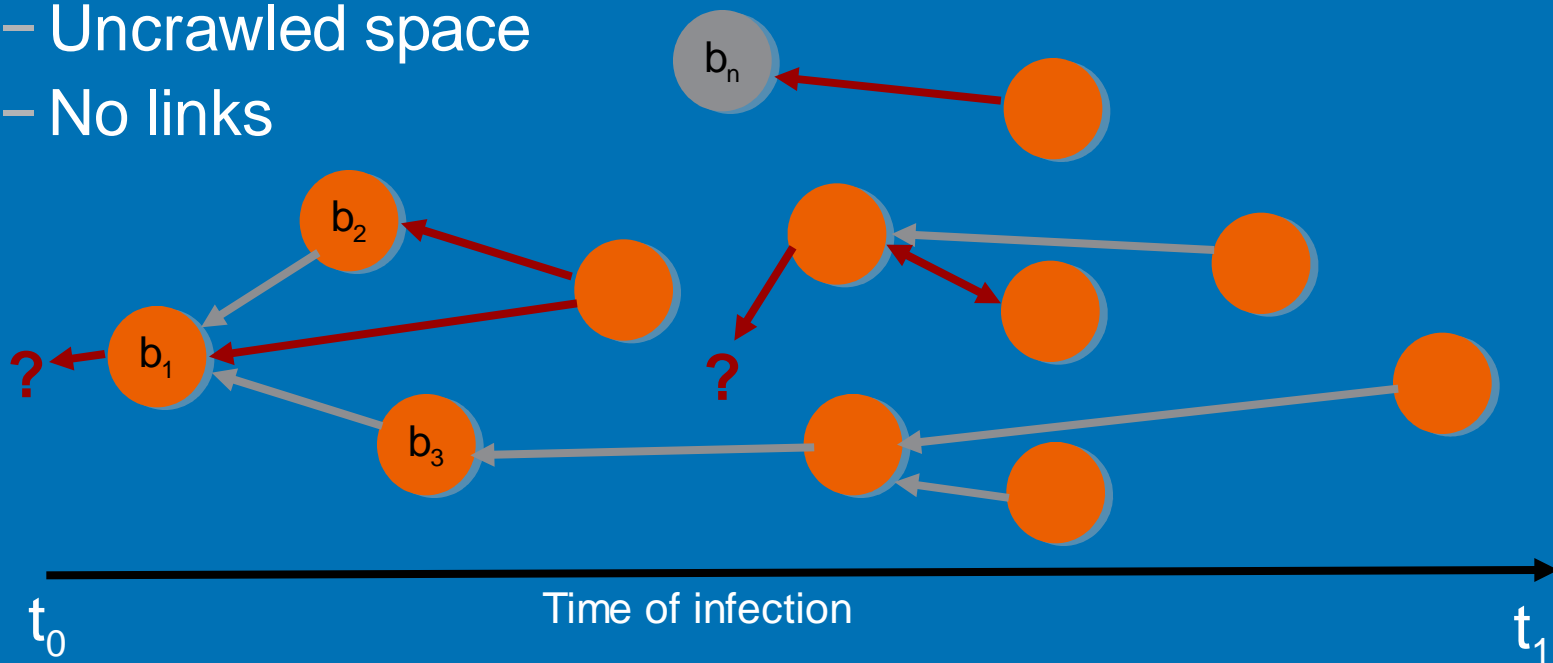
# Microscale Dynamics

- What do we need to talk about specific epidemics?
  - Timings
  - Graphs



# Microscale Dynamics

- Problems
  - Root
  - Multiple possibilities
  - Uncrawled space
  - No links



# Microscale Dynamics

- Easy: Explicit
  - Via links are even better
- Hard: Implicit/Inferred (link inference problem)
  - Use ML algorithm
    - Support Vector Machine (SVM)
    - Logistic Regression
  - What we have going for us
    - Full text
    - Blogs in common
    - Links in common
    - History of infection

# Similarity Measures

- Blogs/Links in Common

$$s(A, B) = n_{AB} / \sqrt{n_A} / \sqrt{n_B}$$

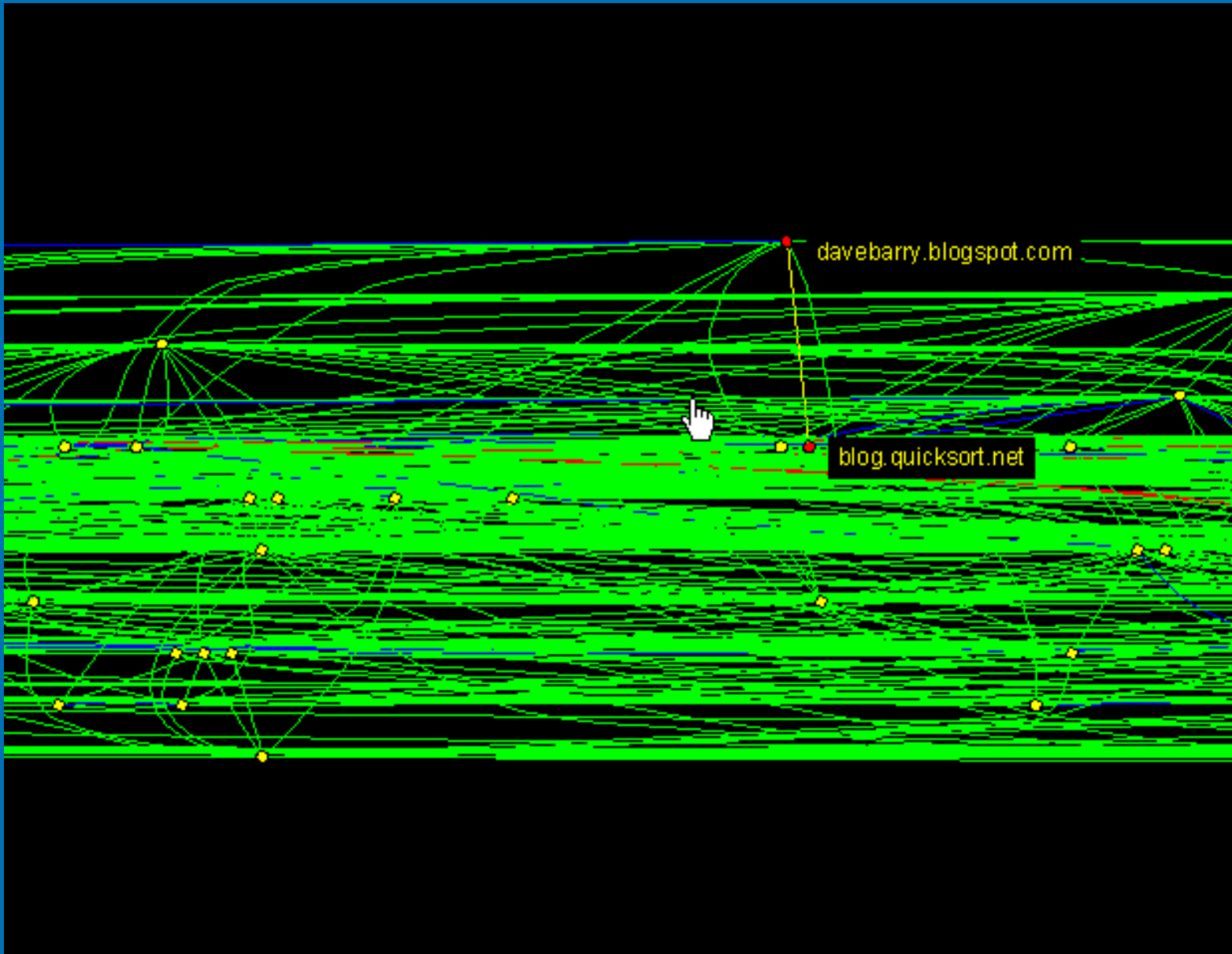
- Text Similarity
  - TFIDF transformation on term vector, cosine metric
- Infection timings
  - $n_{\text{AbeforeB}}/n_A$
  - $n_{\text{AafterB}}/n_A$
  - $n_{\text{AsamedayB}}/n_A$
  - $n_{\text{AbeforeB}}/n_B$
  - $n_{\text{AafterB}}/n_B$
  - $n_{\text{AsamedayB}}/n_B$

# Experiment 1

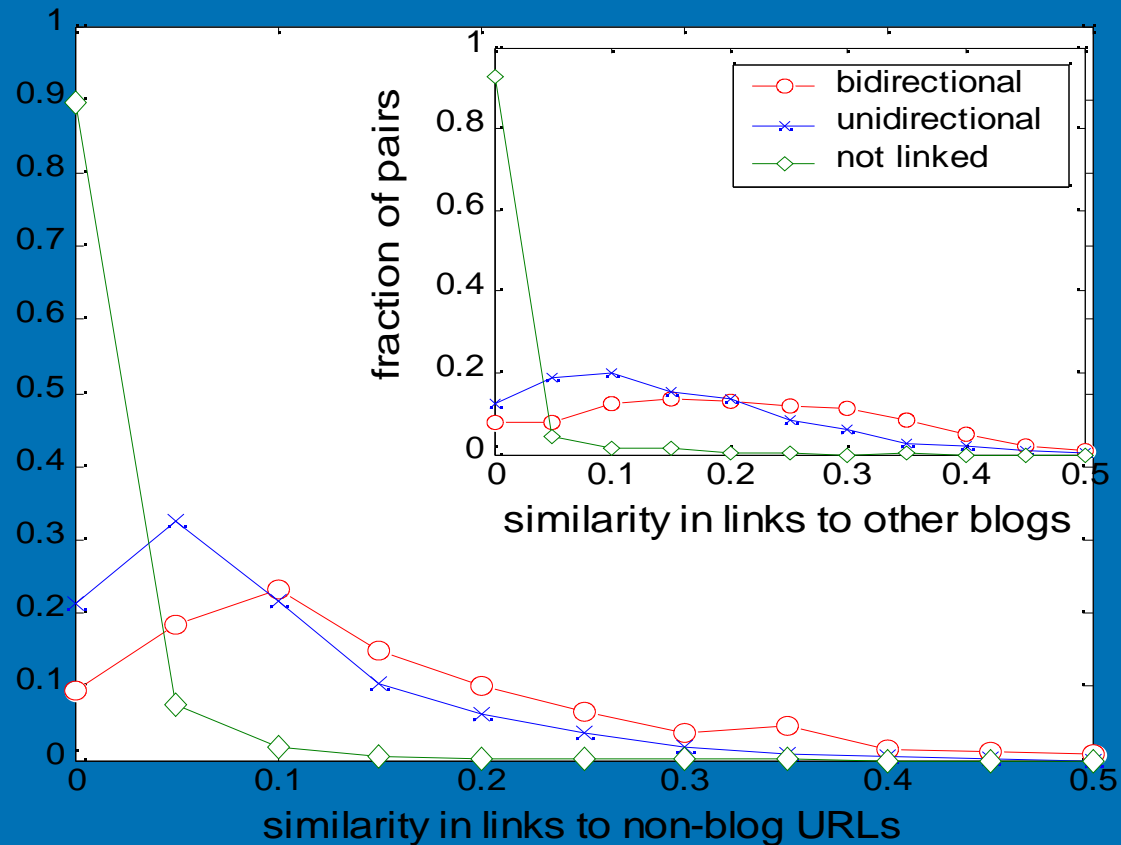
- Find pairs of blogs that are “infected” and linked and pairs that are infected and unlinked
- Train/test on those
- Both classifiers score very well >90%
  
- Problem: doesn't work in the specific case
  - For a given epidemic it connects all blogs

# Experiment 1

- Find pairs of blogs that are “infected” and linked and pairs that are infected and unlinked
- Train/test on those
- Both classifiers score very well >90%
  
- Problem: doesn't work in the specific case
  - For a given epidemic it connects all blogs



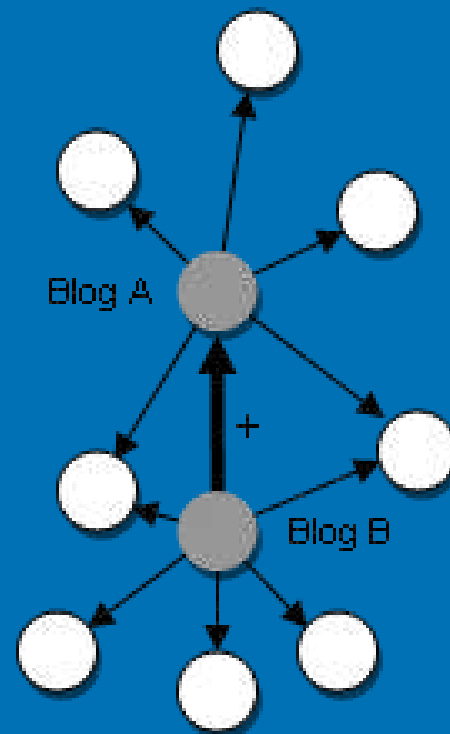
# Link probabilities



- Single shared link = high probability
- Simple classifier (share any): 88%

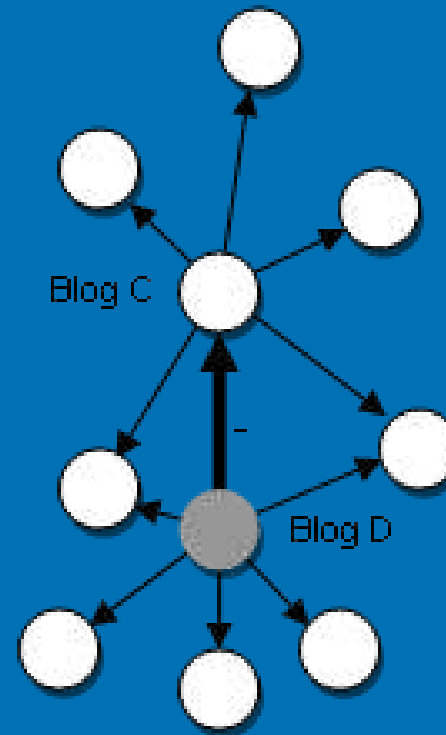
# Experiment 2

**Positive Example**

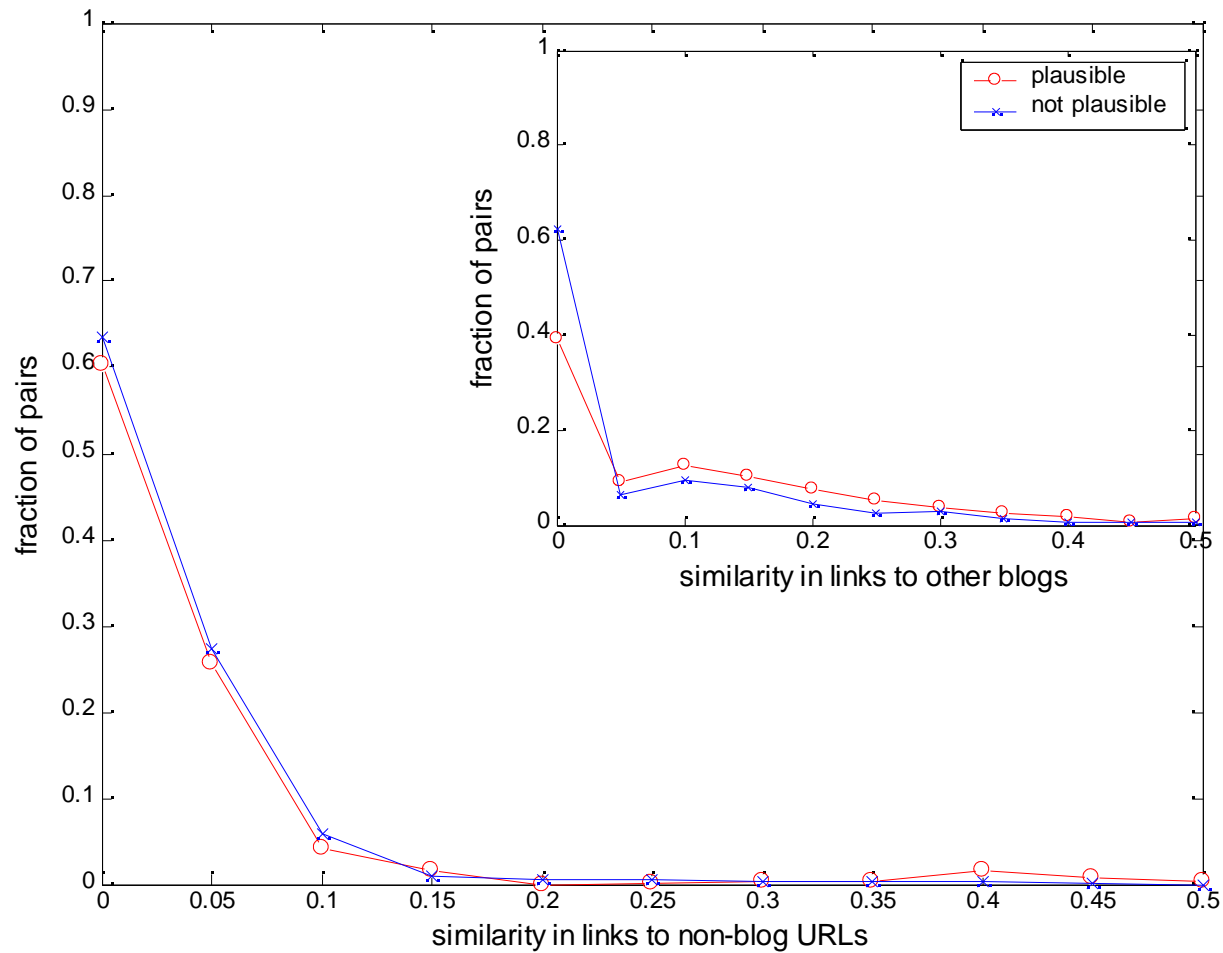


$$T_{incluster}(\text{Blog B}) > T_{incluster}(\text{Blog A})$$

**Negative Example**



# Link Probabilities



# Better & Worse

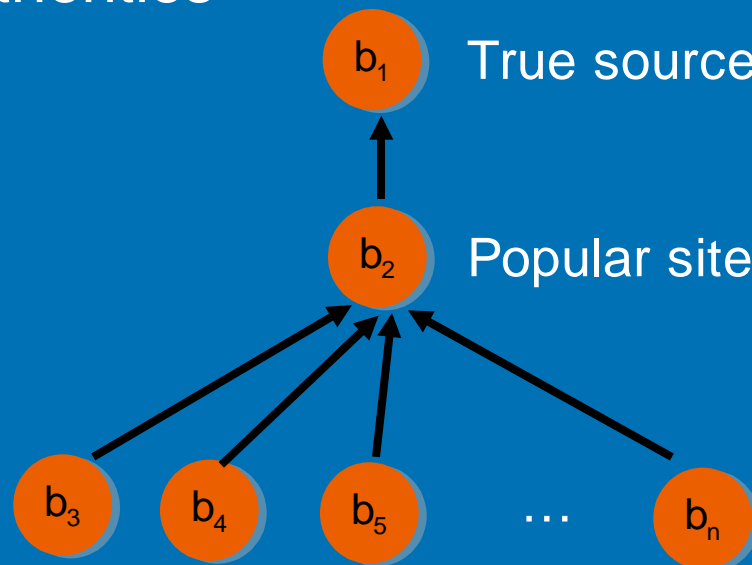
- Doesn't link everything, but...
- Classifier now at ~ 75%
- Hopefully will be better with more data

# Visualization

- Zoomgraph tool
  - Using GraphViz (by AT&T) layouts
- Simple algorithm
  - If single, explicit link exists, draw it
  - Otherwise use ML algorithm
    - Pick the most likely explicit link
    - Pick the most likely possible link
- Tool lets you zoom around space, control threshold, link types, etc.
- Demo...

# iRank

- “Practical” uses of inferred epidemic information
  - Can use a simpler inference (timing)
- Finding good sources
  - Invisible authorities



# Algorithm

- Draw a weighted edge for all pairs of blogs that cite URL  $u$

$$w_{ji}^u = w(\Delta d_{ji})$$

$\Delta d_{ji}$	0	1	2	3	4	5	6	7	>7
$W(\Delta d_{ji})$	2	7	6	5	4	3	2	1	0

- Normalize so outgoing weights = 1
- Do PageRank

# Spamming

- Starting to be an issue for everyone
  - Comment spamming
  - Duplicating whole blog networks
  - Duplicating specific pages
- We require a certain popularity
- Down-weight people who are too effective
- Can also cluster very similar pages into one “meta-blog”

**Table 3. PageRank ordering of blogs**

Rank	Web Site	Rating
1	boingboing.net	166
2	penny-arcade.com	166
3	caoine.org	151
4	slashdot.org	150
5	andrewsullivan.com	117
6	perversiontracker.com	114
7	crazyapplerumors.com	107
8	bloghop.com	106
9	livejournal.com	101
10	dear_raed.blogspot.com	94
11	girlsarepretty.com	85
12	fark.com	83
13	cyberlaw.stanford.edu	80
14	alwayson-network.com	76
15	oddtodd.com	74
16	instapundit.com	70
17	drudgereport.com	70
18	metafilter.com	68
19	wilwheaton.net	64
20	altnet.org	61

**Table 4. iRank ordering of blogs**

Rank	Web Site	Rating
1	blogosphere.us/trends.php	40
2	blogdex.media.mit.edu	32
3	blogosphere.us/trends.php?type=hours	25
4	www2.meeka.dyndns.org:81/~alyssa/	21
5	heliopod.org	21
6	indrasweb.com/blog	18
7	weinstein.org	18
8	pontobr.org	18
9	peiblog.psychoblogger.com/weblog.php	18
10	knoxgeek.com	18
11	ctdata.com	17
12	mosa.unity.ncsu.edu/brabec	17
13	timbu.org/mtblog	16
14	khader.net	15
15	burngreave.net	14
16	forum.b0rken.dk/drupal	13
17	vazdot.info	12
18	nanodot.org	12
19	opencontentlist.com	12
20	inhale.org	12

# Future Work...

- Tie together graph structure to epidemic profile
- Improve inference techniques
- Better evaluation of iRank
  - iRank search engine demo



# Zoomgraph Screenshot

