D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins: **Information Diffusion Through Blogspace**

Mari-Sanna Paukkeri

November 7, 2007

Outline

Introduction
- Related Work
- Corpus Details

Topic Models
- Topic Identification
- Topic Structures: Chatter and Spikes

Modelling Individuals
- Characterizing Individuals
- Individual Propagation
- Transmission Graph
- Extensions

Evaluation & Discussion
- Validation
- Future Work

Introduction

Flow of information in societies affects

- the structure of societies
- relations between different societies

The paper studies:

- Information propagation through a network (blogspace)
  - topic structures and distribution
  - topic propagation from individual to individual
- Analysis based on the *text* of the blog (not hyperlinks)

Information Diffusion

Topics of blog postings

- Identify the topics in the blog data
- Identify the postings that are about each topic
- Characterize how much the topic is *chatter* or *spikes*

Individuals affecting the others

- Define 4 categories of posting behaviour
- Create model for information diffusion
- Learn the parameters of the model from real data
- Identify individuals contributing to "infectious" topics

Related Work

Rich literature around propagation through networks:

- Information Propagation and Epidemics (spread of diseases)
- The Diffusion of Innovation
- Game-Theoretic Approaches

Information Propagation and Epidemics

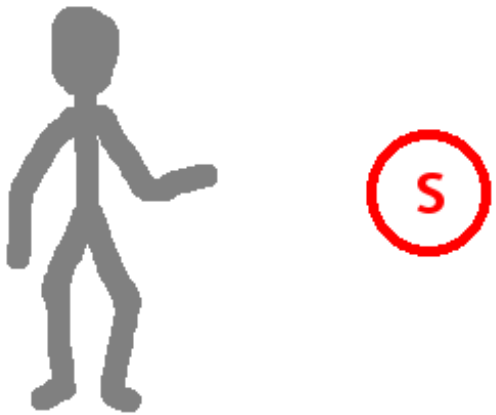
Disease-propagation models (SIR) from epidemiology

- susceptible (S)
- infected/infectious (I)
- recovered/removed (R)

Information Propagation and Epidemics

Disease-propagation models (SIR) from epidemiology

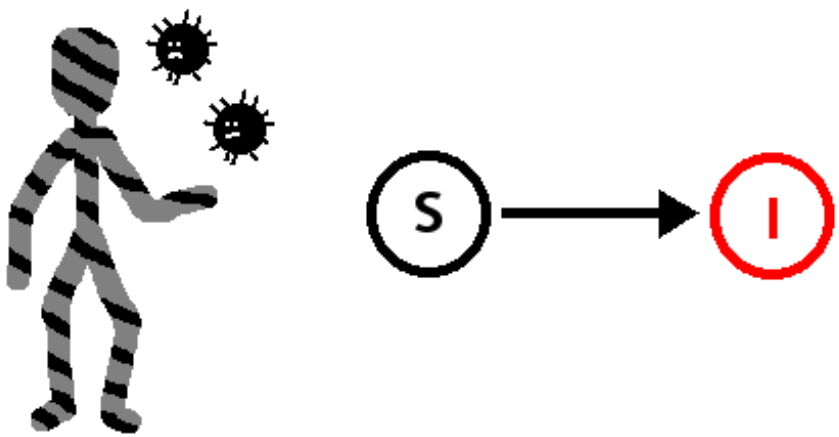
- susceptible (S)
- infected/infectious (I)
- recovered/removed (R)



Information Propagation and Epidemics

Disease-propagation models (SIR) from epidemiology

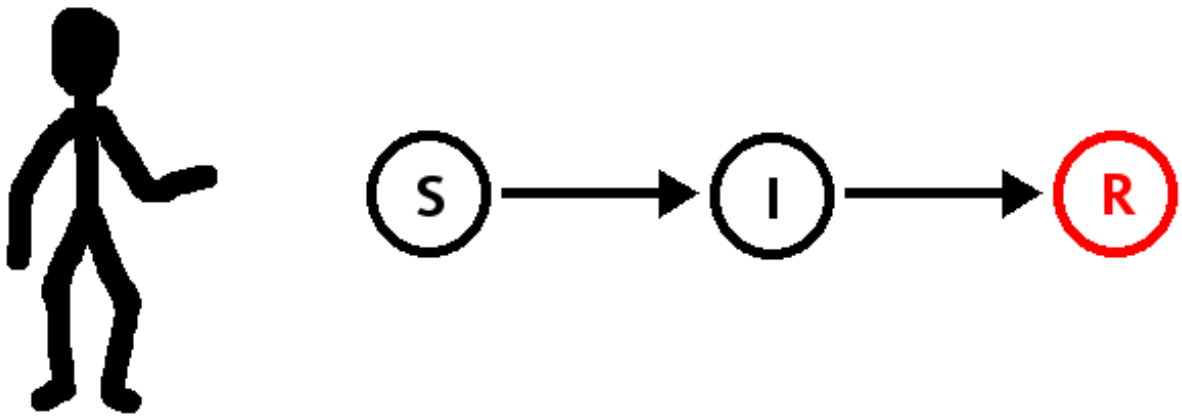
- susceptible (S)
- infected/infectious (I)
- recovered/removed (R)



Information Propagation and Epidemics

Disease-propagation models (SIR) from epidemiology

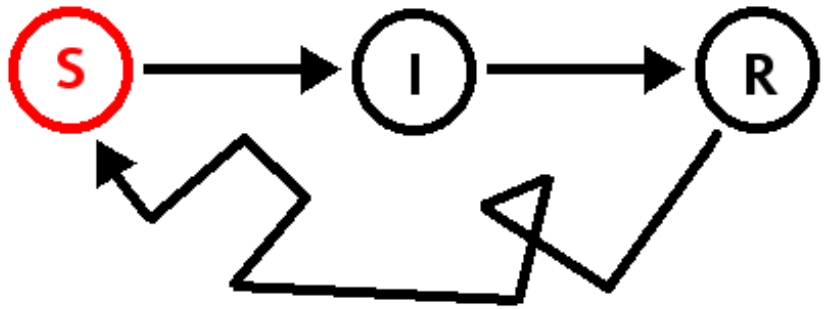
- susceptible (S)
- infected/infectious (I)
- recovered/removed (R)



Information Propagation and Epidemics

Disease-propagation models (SIR) from epidemiology

- susceptible (S)
- infected/infectious (I)
- recovered/removed (R)



Information Propagation and Epidemics

Other epidemic studies:

- SIR model with mutation
- Early studies on fully mixed and homogeneous networks, with random contacts – unrealistic!
- **Epidemic threshold**: the minimum transmission probability for spreading from one seed node to a constant fraction of the entire network
- **Epidemic spreading** on networks that follow a power law
  - probability that the degree of a node is k is proportional to $k^{-\alpha}$.
  - a property of many real-world networks
  - exhibit extremely high error tolerance
- SIS model
- Clustering coefficient
- Transmission model

The Diffusion of Innovation in Social Networks

Modelling adoption of new ideas

- **Threshold models**: each node u has
  - threshold $t_u \in [0, 1]$
  - connection weights $w_{u,v}$
- **Cascade models**
  - If a close node ('friend') of u adopts, there is a chance that u will decide to adopt as well.
- Independent Cascade model
  - directed graph, edges labelled with a probability
  - initially, a non-empty set of nodes is activated
  - at each step, some set of nodes become activated
- General Cascade model
  - generalizes previous: no independence assumption
  - marketing motivation: find the k seed nodes that maximize the expected number of adopters

Game-Theoretic Approaches

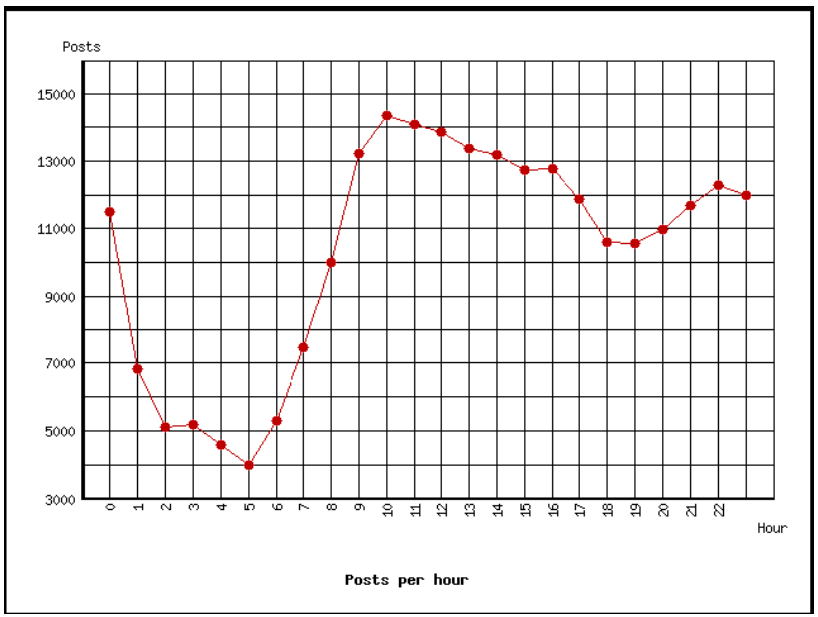
Increase in utility if a player adopts an innovation

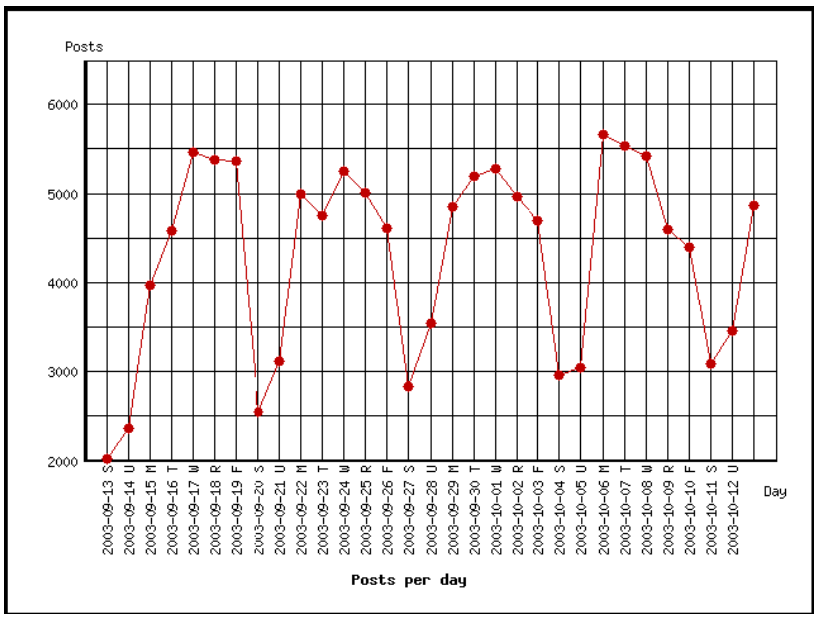
- **Co-ordination game**: in every time step
  - A player chooses a type $\{0, 1\}$, (i.e. 'meme adopted' or not)
  - The player gets payoff for each neighbours having the same type
- Models where agent can selfishly decide whether to form a link or not
  - *cost* of establishing a link
  - *profit* from new information

Corpus

11 804 RSS blog feeds

- 2 000–10 000 blog postings per day
- total of 401 021 postings





14 news channels from `rss.news.yahoo.com`

- identify topics from the major media or real-world events
- crawled hourly

Topic Models

- Topic Identification
- Topic Structures: Chatter and Spikes

Topic Models

Two families of models

- *horizon* models: long-term changes
- *snapshot* models: short-term behaviour

Topic Identification

What is needed?

- A number of important topics
  - different levels: very focused to very broad
- Representative sample of all classes(?) of topics

How is it obtained?

- Proper nouns (11 000 words)
  - all repeated sequences of uppercase words surrounded by lowercase text
- Interesting terms (10 000 words)
  - cumulative inverse document frequency

$$tfcidf(i) = (i-1)\frac{tf(i)}{\sum_{j=0}^{i-1} tf(j)}$$

Chatter and Spikes

A topic is a composition of *chatter* and *spikes*.

- Chatter
  - "background noise"
  - topics of continuous interest
  - topic propagation from blog to blog
- Spike
  - temporary increase in the number of postings on a topic
  - triggered by an event in the Real World - not by another blog posting
- Resonance
  - a posting to which everyone reacts sharply $\longrightarrow$ spike
  - no external input
  - Example:

Chatter and Spikes

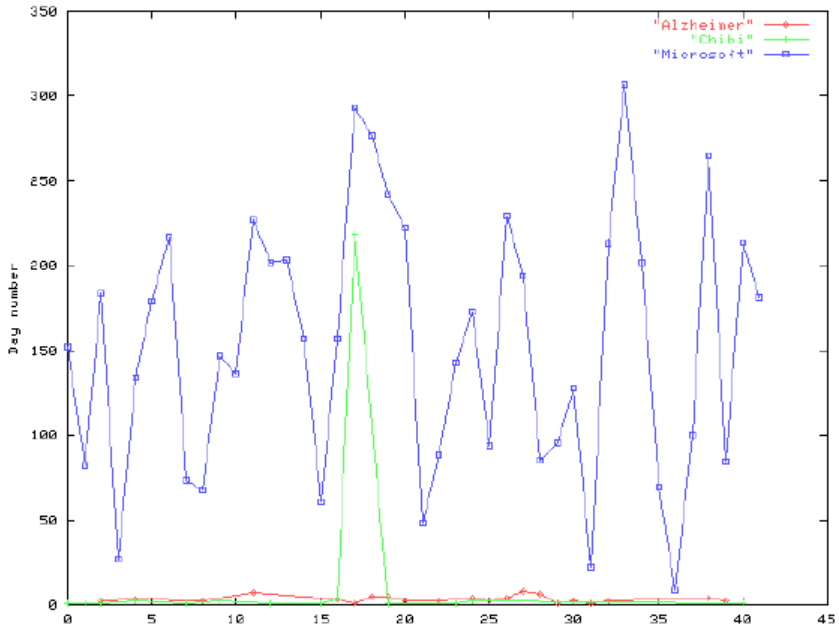
A topic is a composition of *chatter* and *spikes*.

- Chatter
  - "background noise"
  - topics of continuous interest
  - topic propagation from blog to blog
- Spike
  - temporary increase in the number of postings on a topic
  - triggered by an event in the Real World - not by another blog posting
- Resonance
  - a posting to which everyone reacts sharply $\longrightarrow$ spike
  - no external input
  - Example:
    aoccdrnig to rscheearch at an elingsh uinervtisy it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer is at the rghit pclae

Topic = Chatter + Spikes

Categories of topics (depending on the average chatter level and pertinence to the real world)

- **Just Spike** e.g. Chibi (Japanese, means 'dwarf' or 'small child')
- **Spiky Chatter** e.g. Microsoft
- **Mostly Chatter** e.g. Alzheimer's



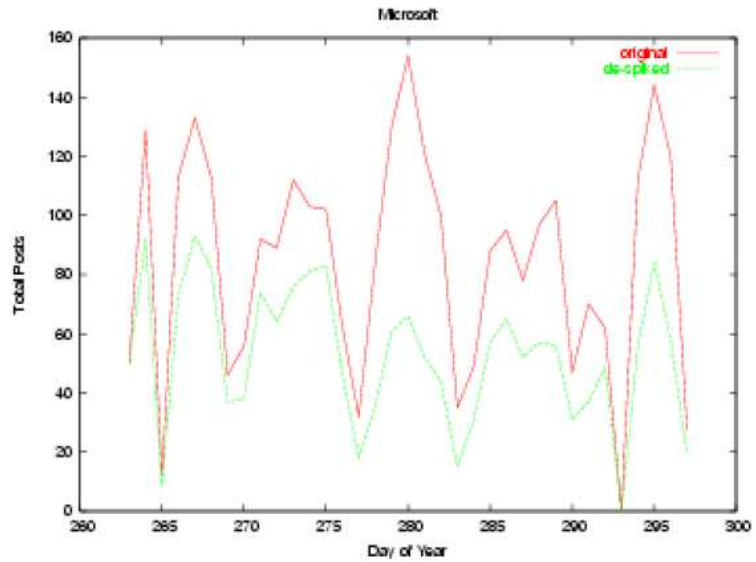
Topic = Chatter + Spiky Subtopics

Identify **subtopics** (case: 'Microsoft')

- select proper nouns x co-occurring with 'Microsoft'
- for each x, compute support s and reverse confidence $c_r = P(target|x)$
- thresholds for s and c_r are 'found'

windows	server	services	longhorn
exchange	ie	office	msdn
outlook	msn	gates	redmond
eolas	xp	netscape	powerpoint
scoble	pdc	motorola	avalon
ms	vb	acrobat	xaml

Top-coverage terms

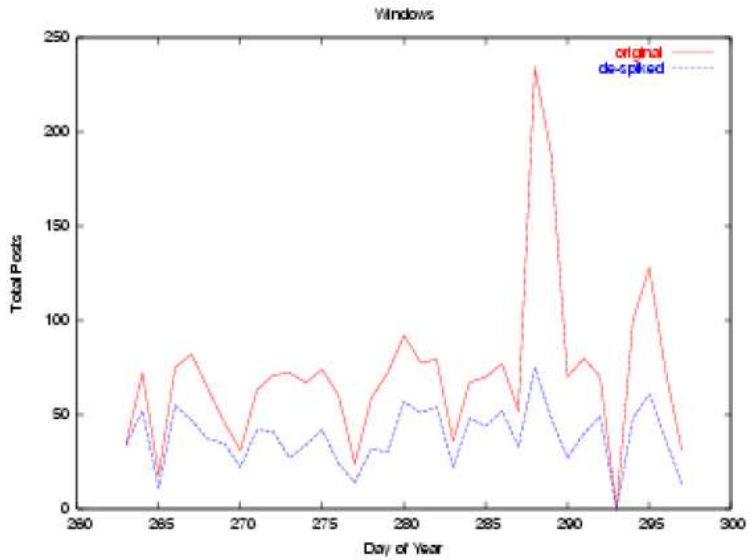


Topic = Chatter + Spiky Subtopics

- Decomposition of subtopic 'Windows'

series	server	os	longhorn
pc	ie	mac	gui
apple	jobs	dell	ui
ram	xp	explorer	drm
unix	pcs	linux	apples
ms	macs	quicktime	macintosh

Top-coverage terms



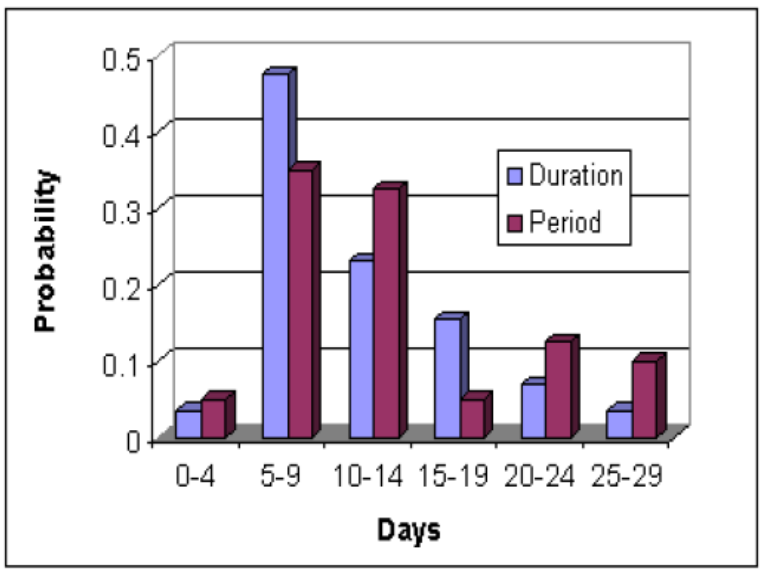
- This study strongly(?) supports the notion of a spike and chatter model of blog posting

Quantitative Characterization of Spikes

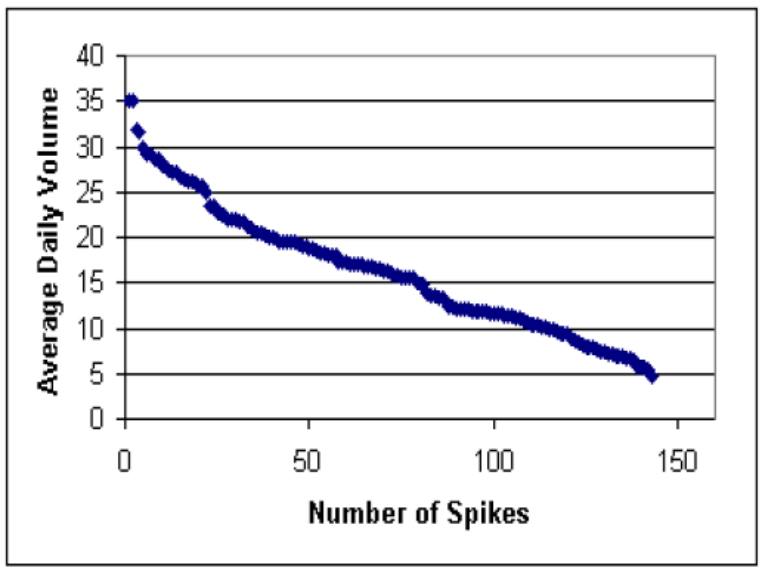
- Average number of posts per day for non-spike regions: 1.6-106. The distribution is well-approximated by

$$Pr[\text{average number of posts per day} > x] \sim ce^{-x}$$

- Duration and period of spikes



- Average daily volume for spike periods



Modelling Individuals

- Characterizing Individuals
- Individual Propagation
- Transmission Graph
- Extensions

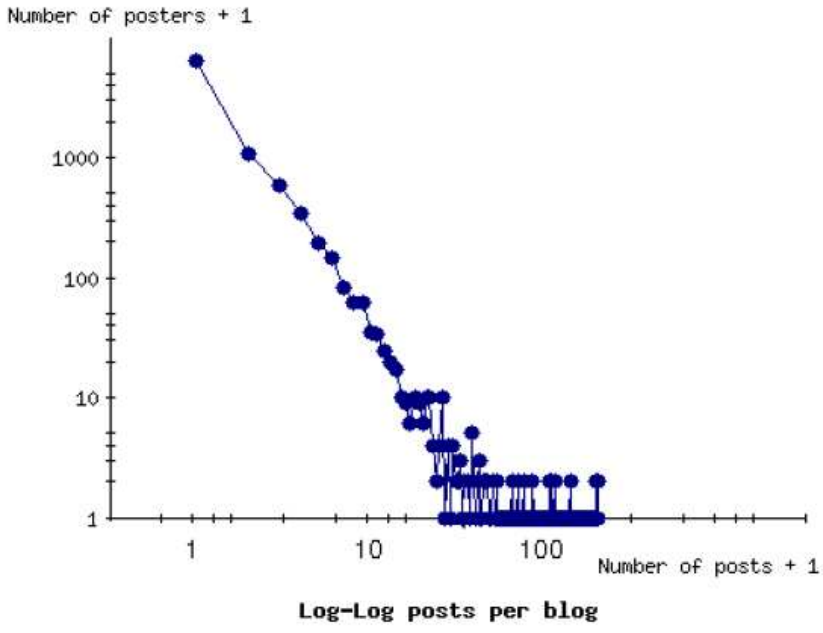
Modelling the Path of Topics Through Individuals

Model individuals affecting the others

- Define 4 categories of posting behaviour
- Create model for information diffusion through blogspace
- Learn the parameters of the model from real data
- Identify individuals contributing to "infectious" topics
- A directed graph

Characterizing Individuals

- Number of posts per user in the data-collection window



- Classify these to following classes:
  - Ramp-Up - First 20% of post mass
  - Ramp-Down - Last 20%
  - Mid-High - Middle 25%
  - Spike

Individual Propagation

- SIR model (no multiple postings on a topic)
- A user may visit certain blogs frequently, and other blogs infrequently
  - edge parameter $r_{u,v}$: probability that u reads v's blog any given day
- Information propagation: probability that the topic will propagate from u to neighbouring v
  - u reads v's post with *reading probability* $r_{u,v}$
  - δ - a delay from exponential distribution with parameter $r_{u,v}$
  - v chooses to write about the same topic with probability $\kappa_{u,v}$

Induction of the Transmission Graph

- *Closed-world assumption*: all occurrences of a topic except the first are the result of communication via edges in the network
- Select topic (URL, phrase, name, . . .)
- Gather blog entries containing the topic into a list
- Sort list by publication date $\longrightarrow$ *traversal sequence*
- EM-like algorithm to induce the parameters of the transmission graph
  - Initialize r and κ
  - **Soft assignment step**: compute for each topic and each pair (u, v) the probability that the topic traversed the (u, v) edge
  - **Parameter-Update step**: For fixed u and v, recompute r (reading probability) and κ (probability to copy the topic)

Algorithm

- r reading probability
- κ copy probability
- δ delay (in days) between u and v

Soft-Assignment Step

$$p_{u,v} = \frac{r(1-r)^{\delta}\kappa}{\sum_{w<v} r_{w,v}(1-r_{w,v})^{\delta_{w,v}}\kappa_{w,v}}$$

Parameter-Update Step

$$r := \frac{\sum_{j\in S_1} p_j}{\sum_{j\in S_1} p_j \delta_j} \qquad \kappa := \frac{\sum_{j\in S_1} p_j}{\sum_{j\in S_1\cup S_2} \Pr[r \le \delta_j]}$$

Extensions to the Model

- The Real World
  - the topic might be read both from major media and from other blogs
- Span of attention
  - people do not have time to read all the blogs
    $\longrightarrow$ limit in-degree of nodes (e.g. *attention threshold* parameter)
- Stickiness
  - certain topics are more interesting than others
    $\longrightarrow$ another parameter *stickiness*
- Multiple posts
  - authors routinely write multiple posts on the same topic

Evaluation & Discussion

- Validation
- Future Work

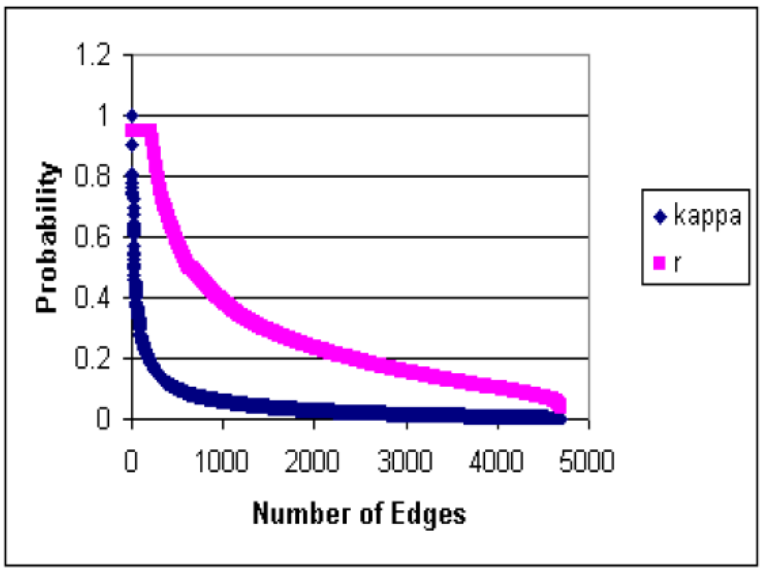
Validation

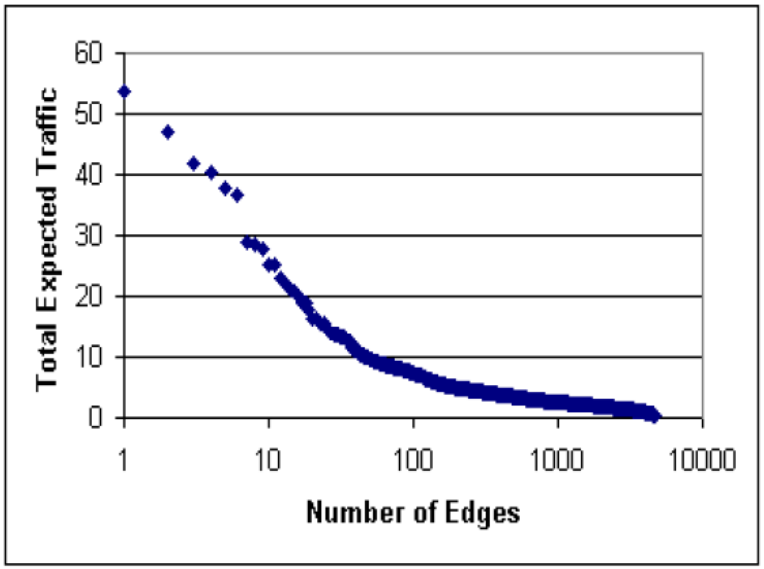
Synthetic data:

- Series of propagation networks

Real data:

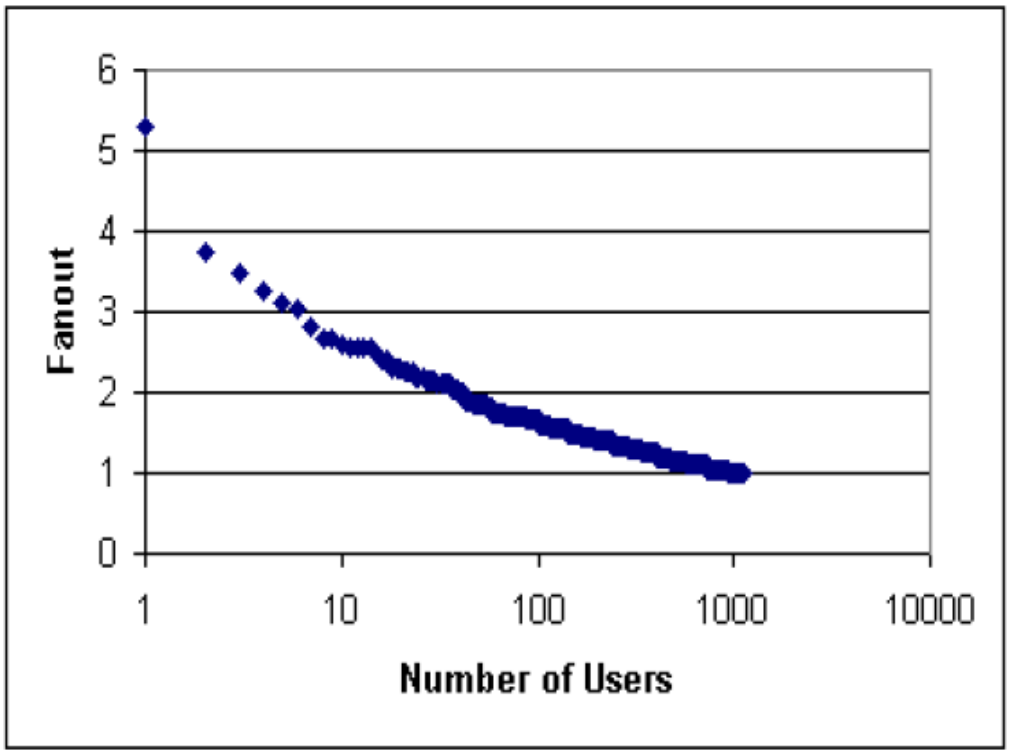
- 100 blogs from `http://blogstreet.com`
- 70 of them were in the RSS-generated dataset





Fanout

- Certain individuals are likely to pass topics on to many friends
- **Fanout**: Expected number of follow-on infections generated by each person



Future Work and Applications

News services

- The large volume of blogs makes it difficult to identify the crucial posts in high-chatter topics (same with corporate press releases)
- The proposed model enables identification of subtopics that are experiencing spikes

Marketing

- Weblogs offer and inexpensive and nearly real-time tool for evaluating the effectiveness of company's image-affecting activities (say, advertising)