Social Networks: Applications

Advanced Computer Networks

Summer Semester 2013

Recap: Information Cascading Model

- Consider an urn with 3 marbles. It can be either:
  - Majority-blue: 2 blue, 1 red
  - Majority-red: 1 blue, 2 red
- Each person wants to best guess whether the urn is majority-blue or majority-red
- Experiment: making decision one by one
- Analysis: Bayes' Rule (posterior probability)

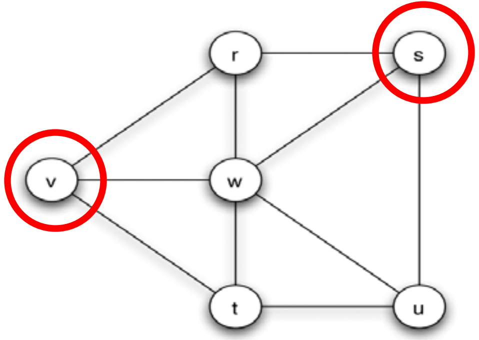
$$\Pr[A \mid B] = \frac{\Pr[A] \cdot \Pr[B \mid A]}{\Pr[B]}.$$

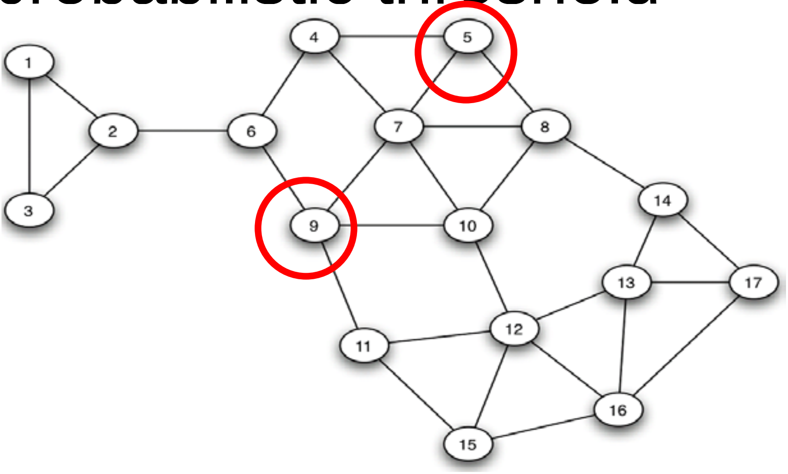
- Decision: based on the best probability

$$Pr[majority - blue|blue, blue, red] = \frac{2}{3} \geq \frac{1}{2}$$

Recap: Social Influence Maximization

- Problem: Finding k most influential nodes in a social network using the probabilistic threshold model

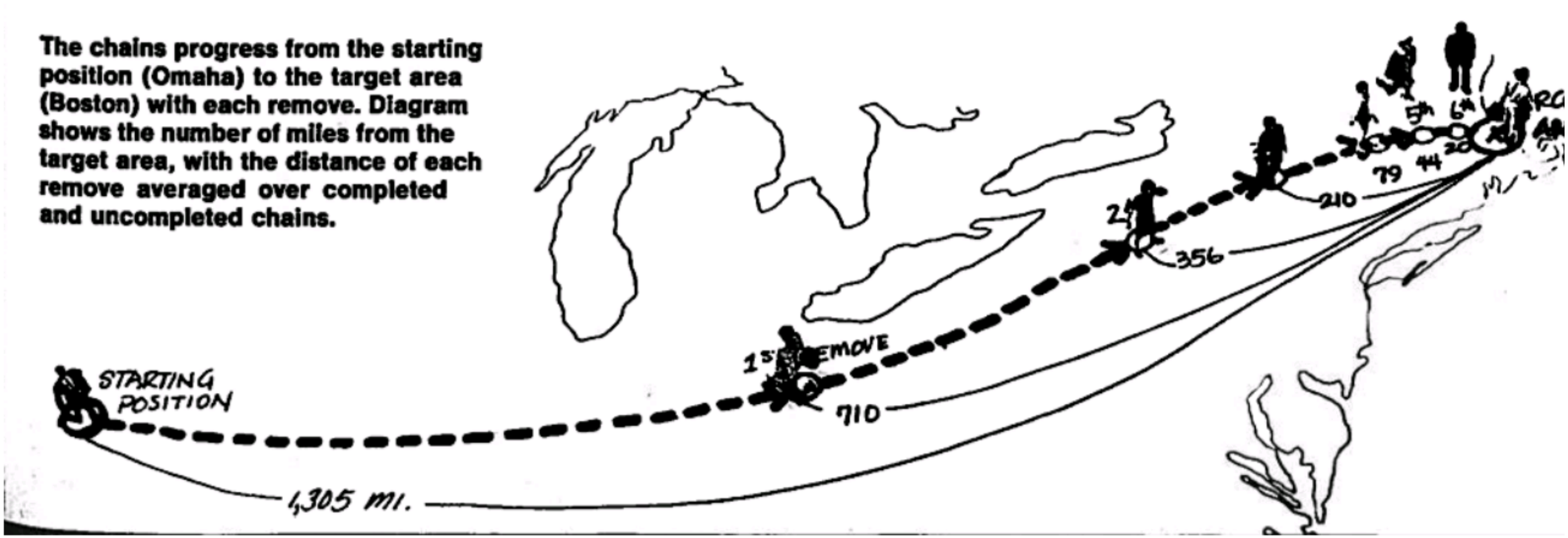




- The problem is shown to be NP-hard
- Heuristic solutions
  - Node degree
  - Distance centrality
  - Greedy

Decentralized Search

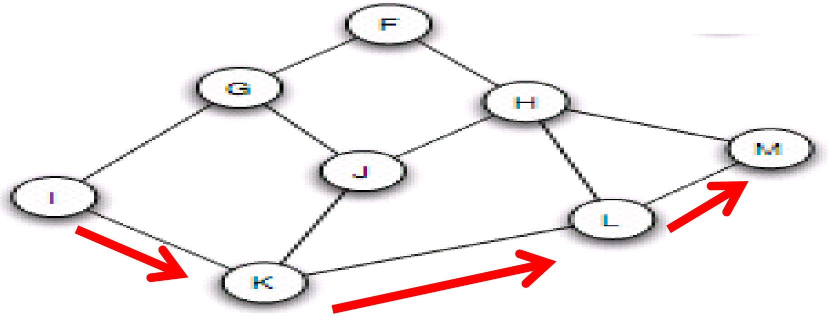
- Question
  - In a Small World network, how to find the short path between a pair of nodes?
    - Centralized strategy?
    - Flooding?
    - Milgram experiment: people collectively find short paths to the designated target -> decentralized search is possible



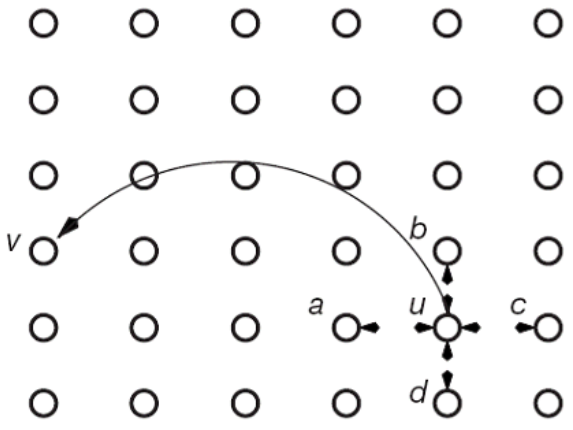
The chains progress from the starting position (Omaha) to the target area (Boston) with each remove. Diagram shows the number of miles from the target area, with the distance of each remove averaged over completed and uncompleted chains.

Decentralized Search

- Node s sending a message to destination t
  - s only knows locations of its friends and locations of the target t
  - s only has local information, it does not know links of other nodes
- Principle: s send the message to its friend who is the closest to t
- Search path length: the number of steps to reach t



A General Network Model



- One dimension: A ring
- Two dimension: A grid
- Each node has only one long link
- The probability of a long link from u to v is:

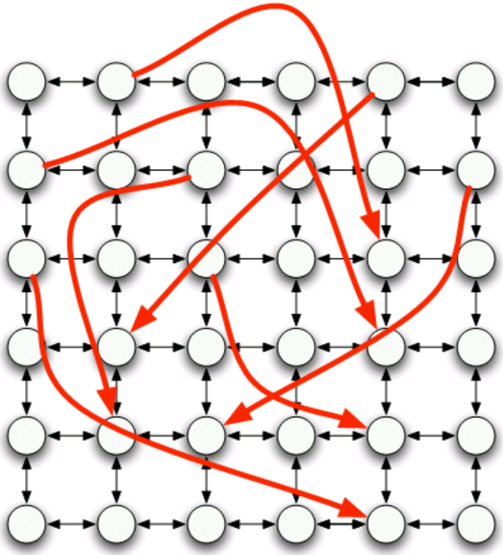
$$Pr\{u \rightarrow v\} \sim d(u,v)^{-q}$$

- Where d(u,v) is the distance (grid steps) between node u and v, and q is a parameter

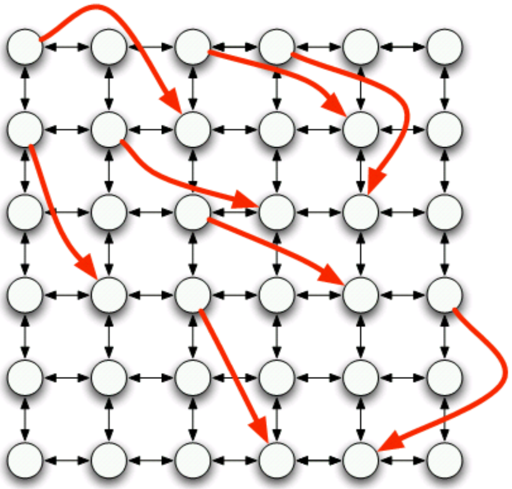
Choosing the parameter q

$$Pr\{u \rightarrow v\} \sim d(u,v)^{-q}$$

- Different q yields different networks, which have different shortest path lengths
- q=0: just like the Watts-Strogatz model
- q-> +∞: only links to nearby nodes



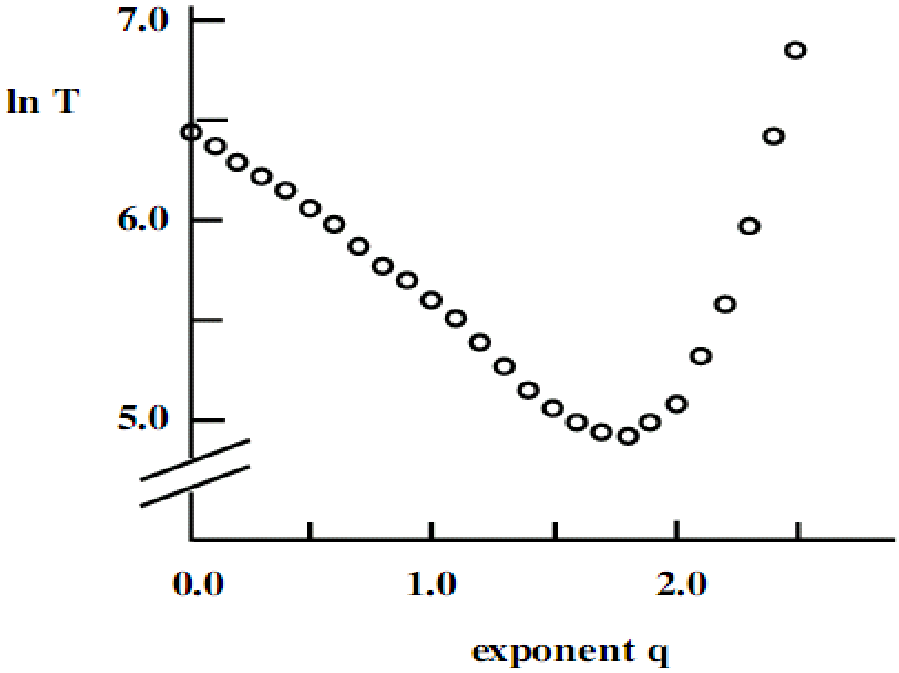
Too random!



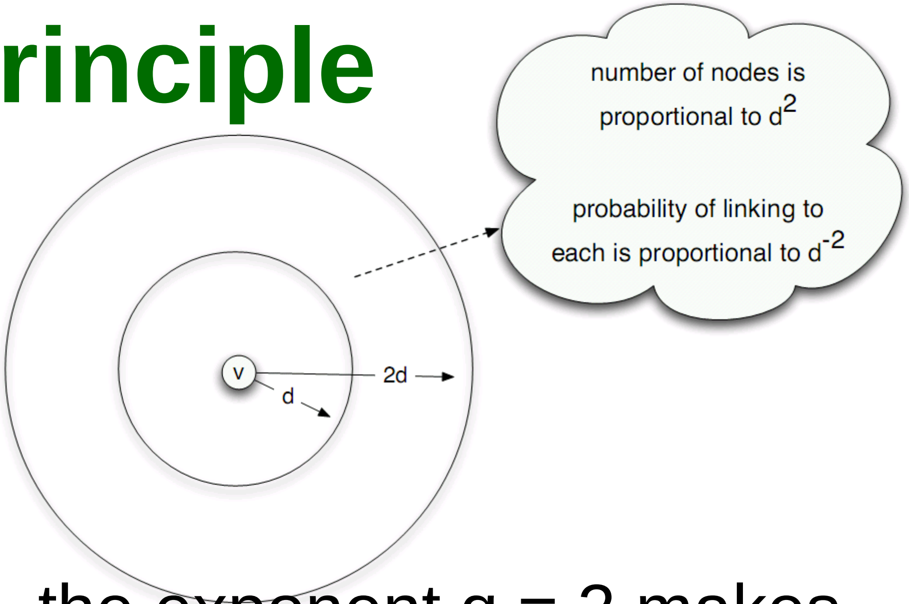
Not random enough!

What is the best value of q?

- Is there a value of q, making the search path achieves the shortest?
- Experiment on a two-dimensional grid
  - q≈2



Inverse-Square Principle



number of nodes is proportional to d^2

probability of linking to each is proportional to d^{-2}

- For a two-dimensional grid, the exponent q = 2 makes it best for decentralized search

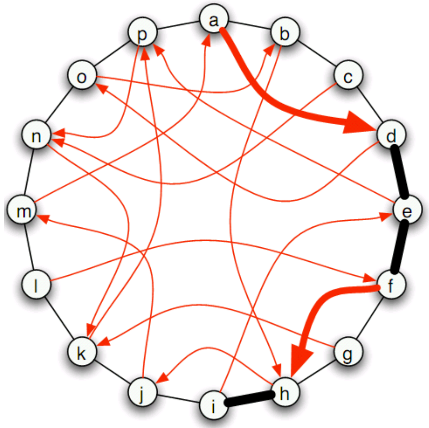
$$Pr\{u \rightarrow v\} \sim d(u,v)^{-2}$$

  - Guess: for d-dimensional, q=d!

- Rough explanation
  - The total number of nodes in an area is proportional to d^2
  - The probability for v linking to the nodes is proportional to d^{-2}
  - They cancel out -> making the probability from v to any other node in the area is independent of d

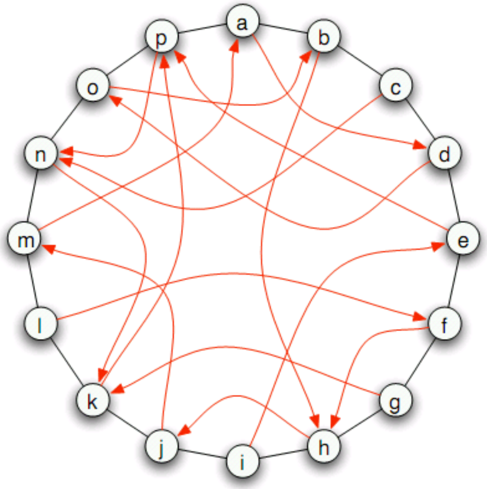
Analysis the Model in 1-dimension

- Nodes are arranged in a ring.
- For 1 dimension, p=1 is the best $Pr\{u \to v\} \sim d(u,v)^{-1}$
- Each node knows only local information, performing decentralized search
- Search strategy: Myopic search
  - When a node v is holding the message, it passes it to the contact that lies as close to t on the ring as possible
  - Not guarantee to be shortest path
- Question: what is the expected length of search path?

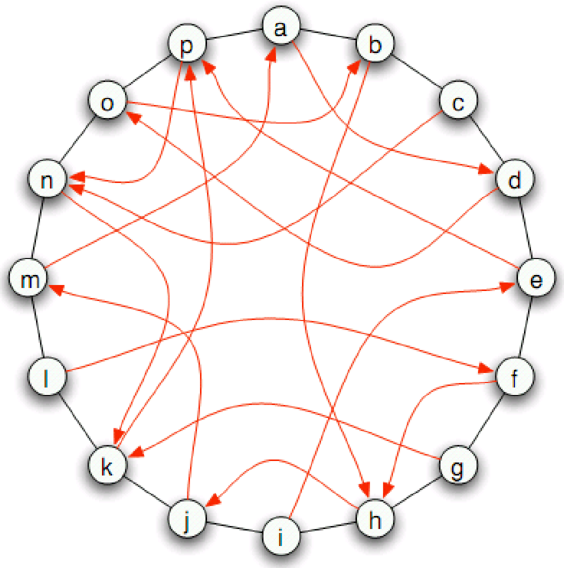


- Claim: for q=1 in 1-dimensional model, we can get from s to t in O(log(n)^2) steps.
- Proof:
- Normalization: $Pr\{u \to v\} \sim d(u,v)^{-1}$
  - Let $$Z = \sum_{i \neq u} d(u,i)^{-1}$$
  - The probability of linking from node u to v is:

$$Pr\{u \to v\} = \frac{d(u,v)^{-1}}{Z}$$



(b) *A ring augmented with random long-range links.*



(b) *A ring augmented with random long-range links.*

- Since

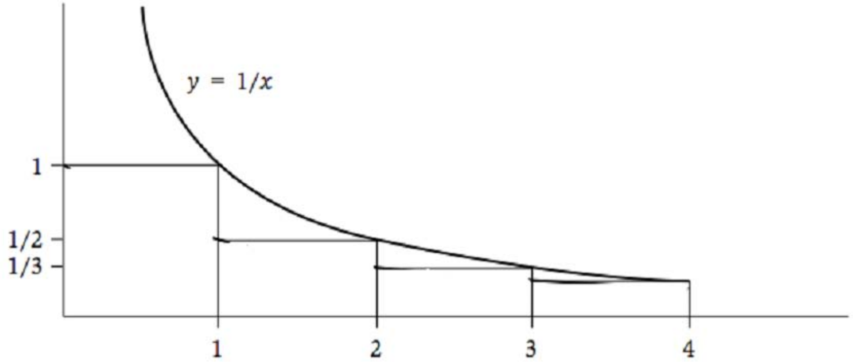
$$Z = \sum_{i \neq u} d(u,i)^{-1}$$

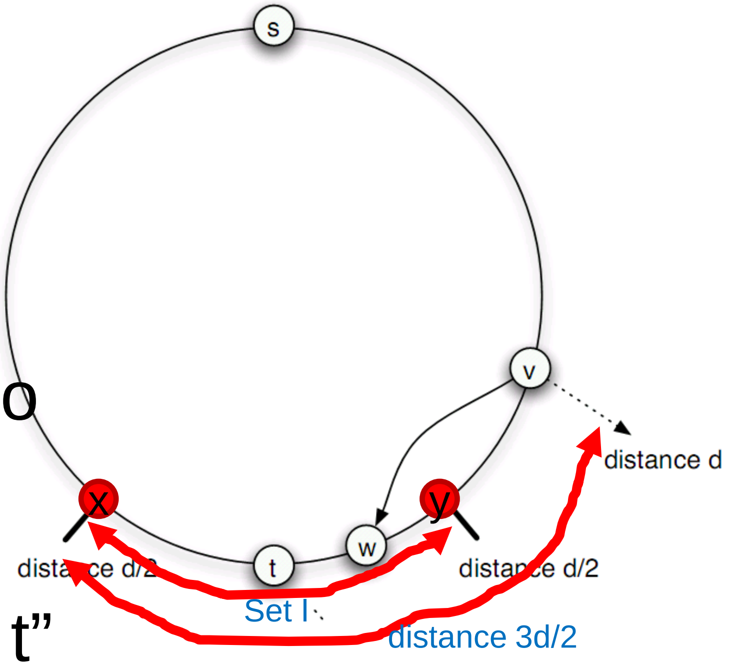
$$Z \leq 2\left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n/2}\right)$$

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{k} \leq 1 + \int_1^k \frac{1}{x} dx = 1 + \ln k.$$

- We have

$$Z \leq 2(1 + ln(n/2)) = 2ln(n)$$



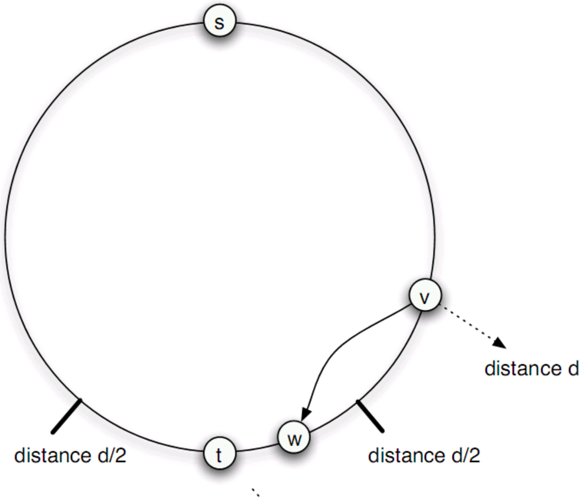


- For a node v, assume its distance to destination t is d, when will the message enters d/2 of t?
- Let I = “the set of nodes with d/2 of t”
  - The number of nodes in I is d+1

$$Pr\{v\ points\ to\ I\} = \sum_{j\in I} Pr\{v \rightarrow j\} = \sum_{j\in I} \frac{d(v,j)^{-1}}{Z}$$

$$= \frac{1}{Z}\sum_{j\in I}\frac{1}{d(v,j)} \geq \frac{1}{Z}(d+1)\frac{1}{d(v,x)} \geq \frac{1}{Z}d\frac{2}{3d} \geq \frac{2}{3Z}$$

- Since $Z \leq 2ln(n)$
- We have $Pr\{v\ points\ to\ I\} \geq \dfrac{1}{3ln(n)}$



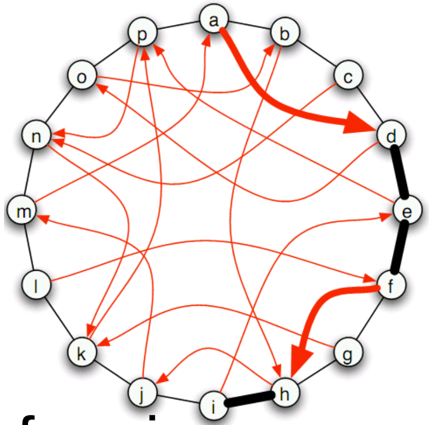
- We have

$$Pr\{v\ points\ to\ I\} \geq \frac{1}{3ln(n)} = O(\frac{1}{ln(n)})$$

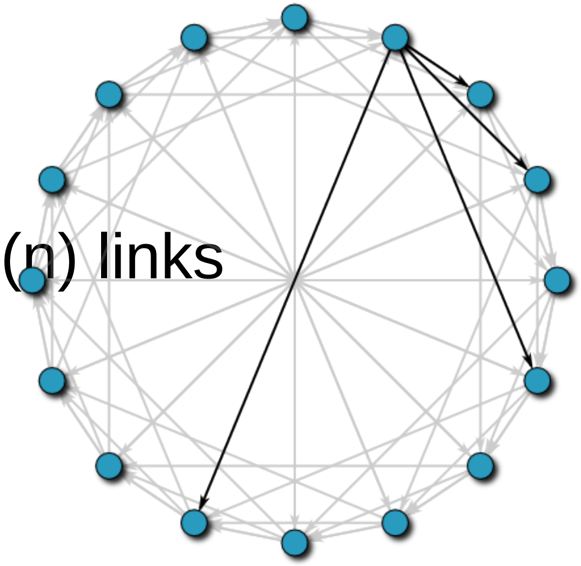
- It means within O(ln(n)) steps, we can get into I from v (the distance is halved!)
- Distance can be halved at most $\log_2$(n) times, so the expected time from s to t is

$$\mathrm{O}(\ln(n) \cdot \log_2(n)) = \mathbf{O}(\mathbf{log}(\boldsymbol{n})^2)$$

Summary



- In 1-dimenstional ring structure
  - Each node knows only local information, performing decentralized search
  - Search strategy: Myopic search
  - p=1 achieves the shortest search path length
  - Expected search path: $O(\log(n)^2)$
- Compare with P2P searching?
  - Chord
  - Each node has a FingerTable with log(n) links
  - The search path length is O(log(n)).



Analysis in Two Dimensions

- For 2-dimensional grid, q=2 achieves the best for decentralized searching
  - For n-dimensional, should be q=n.
- Analysis is similar to 1-demensional case
  - Normalization: z is still O(ln(n)) $Z = \sum_{i \neq u} d(u,i)^{-2}$
  - The number of nodes within d/2 of the target is $O(d^2)$
  - The probability v link to one node in I is $O(1/d^2 Z)$
  - The probability of halving the distance is:

    $O(d^2) * O(1/d^2 Z) = O(1/Z)$ (d is canceled out!)
    - Similar, for n-dimensional, letting q=n will cancel out d
  - The expected steps to halve the distance is O(Z)=O(ln(n))
  - The total expected steps from s to t is: $\log_2 n * O(\ln(n)) = O(\log(n)^2)$
- This is called Inverse-Square Principle $Pr\{u \rightarrow v\} \sim d(u,v)^{-2}$

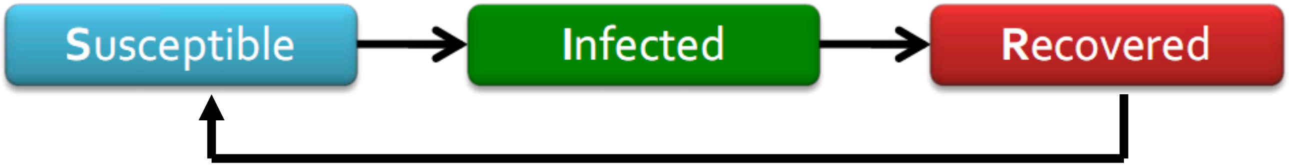
Epidemics

Spread of Contagious Diseases

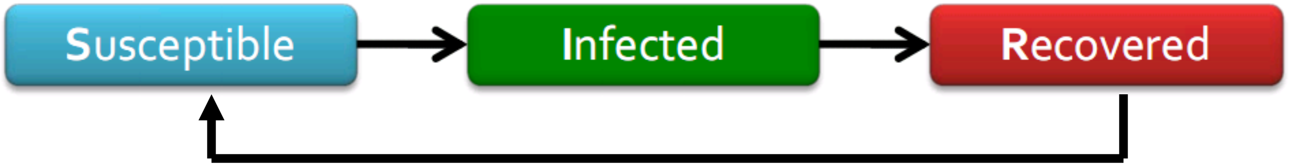
- Spread of contagious diseases
  - Can pass explosively through a population
  - Determined by the properties of the virus: including its contagiousness, the length of its infectious period, and its severity
  - Also affected by network structures within the population it is affecting
- The spread of computer viruses
- Diffusion of ideas through social networks

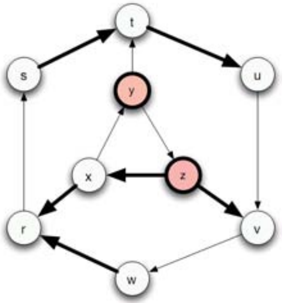
The SIRS Epidemic Model

- Each node goes through the following potential stages: S-I-R-S
  - Susceptible: Before the node has caught the disease, it is susceptible to infection from its neighbors.
  - Infectious: Once the node has caught the disease, it is infectious and has some probability of infecting each of its susceptible neighbors.
  - Removed: After a particular node has experienced the full infectious period, this node is removed from consideration, since it no longer poses a threat of future infection.
  - Susceptible: after the removed stage, it returns to the Susceptible stage



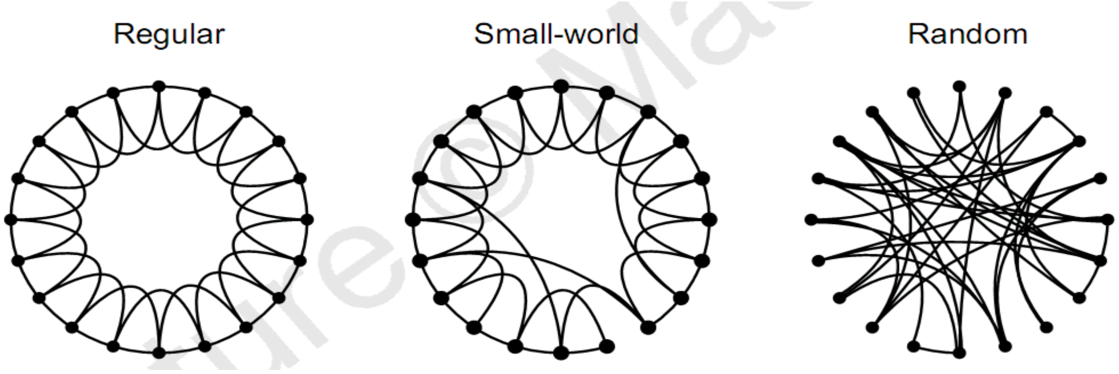
- Process
  - Initially, some nodes are in the I state and all others are in the S state.
  - Each node v that enters the I state remains infectious for a fixed number of steps t_I .
  - During each of these t_I steps, v has a probability p of passing the disease to each of its susceptible neighbors.
  - After t_I steps, node v is no longer infectious. It then enters the R state for a fixed number of steps t_R. During this time, it cannot be infected with the disease, nor does it transmit the disease to other nodes.
  - After t_R steps in the R state, node v returns to the S state.


Susceptible
Infected
Recovered



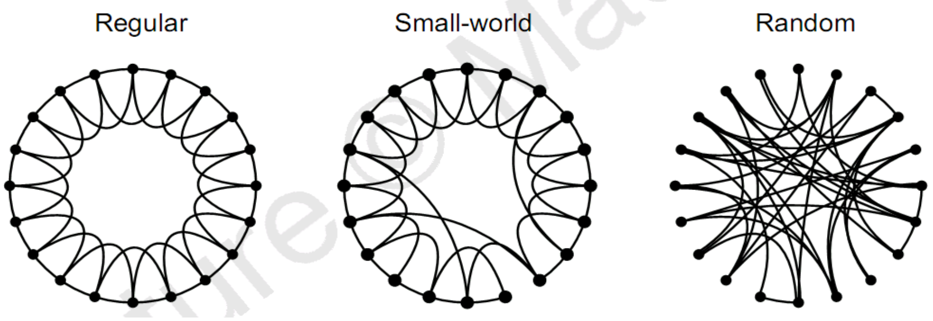
Small-World Contact Networks

- **Recap: The 1-deminsional Watts-Strogatz Model**
- Starting from a ring lattice with n vertices and k edges per vertex.
  - Regular network with high clustering coefficient
- We rewire each edge at random with probability p (0≤p≤1).
  - p=0: regular network
  - p=1: random network
  - 0<p<1: small world network



Small World Effect in the SIRS Model [1]

- Different behavior is observed depending on the value of p
  - When p is small (p=0.01)
    - Disease transmission through the network occurs mainly via the short-range local edges
    - Flare-ups of the disease in one part of the network never become coordinated with other parts
  - When p increases (p=0.2)
    - These flare-ups start to synchronize
    - Oscillations intermittently appear and then disappear
  - For very large values of p (p=0.9)
    - There are clear waves in the number of affected individuals



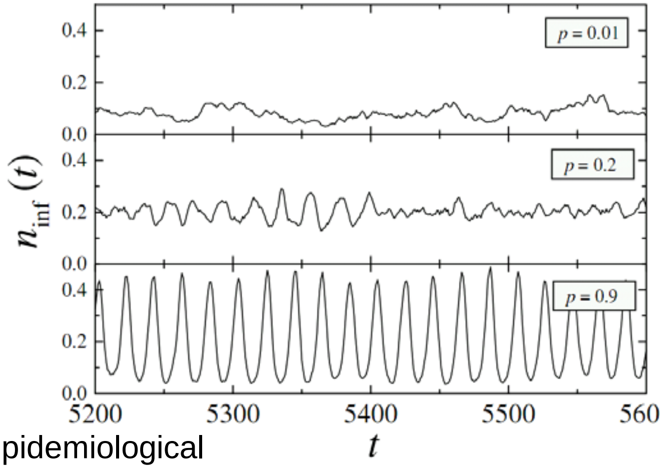


FIG. 1. Fraction of infected elements as a function of time.

[1] Marcelo Kuperman and Guillermo Abramson, Small World Effect in an Epidemiological Model, PHYSICAL REVIEW LETTERS, Vol. 86, No. 13, 2001, 2909-2012.

Lecture next week

- Lecture room is temporary moved to IFI 0.101 for next week (June 13).

Case Study: Tracking Flu Using Twitter [2]

- Collecting information about epidemics:
  - The location, timing and intensity of an epidemic
  - Information is collected from school and workforce absenteeism figures, phone calls and visits to doctors and hospitals
- Gathering this information is a difficult, resource-demanding, time-consuming procedure
- Use of search engine data to detect Influenza-like Illness (ILI)
  - Geographic clusters with a heightened proportion of health-related queries
- Using Twitter to detect ILI?

[2] Vasileios Lampos and Nello Cristianini, Tracking the flu pandemic by monitoring the Social Web, the 2nd International Workshop on Cognitive Information Processing, 2010

Data

- Twitter, UK
  - Daily average of 160,000 tweets
  - 24 weeks from 06/22/2009 to 12/06/2009
  - Twitter geolocation (geographical coordinates).
- Official health reports
  - Health Protection Agency (HPA), UK.
  - Region A = Central England & Wales
  - Region B = South England
  - Region C = North England
  - Region D = England & Wales
  - Region E = Wales & Northern Ireland

Official Health Reports

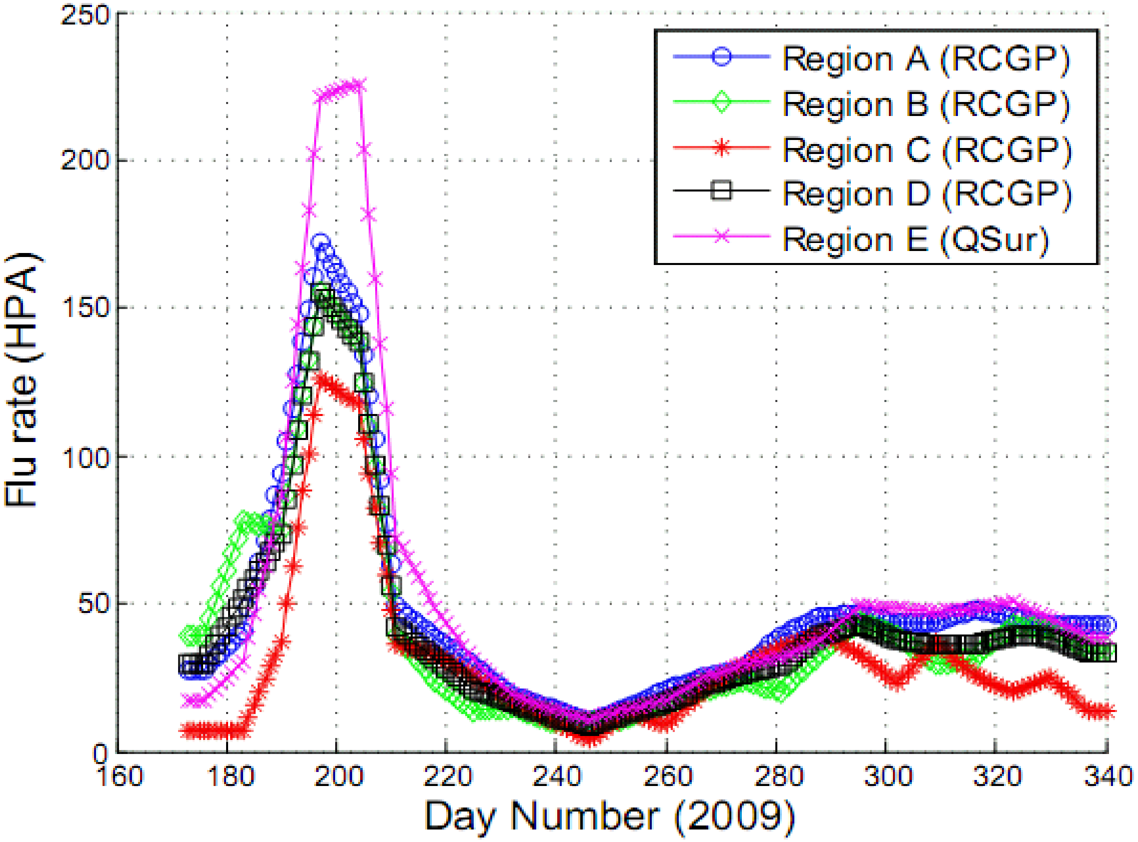
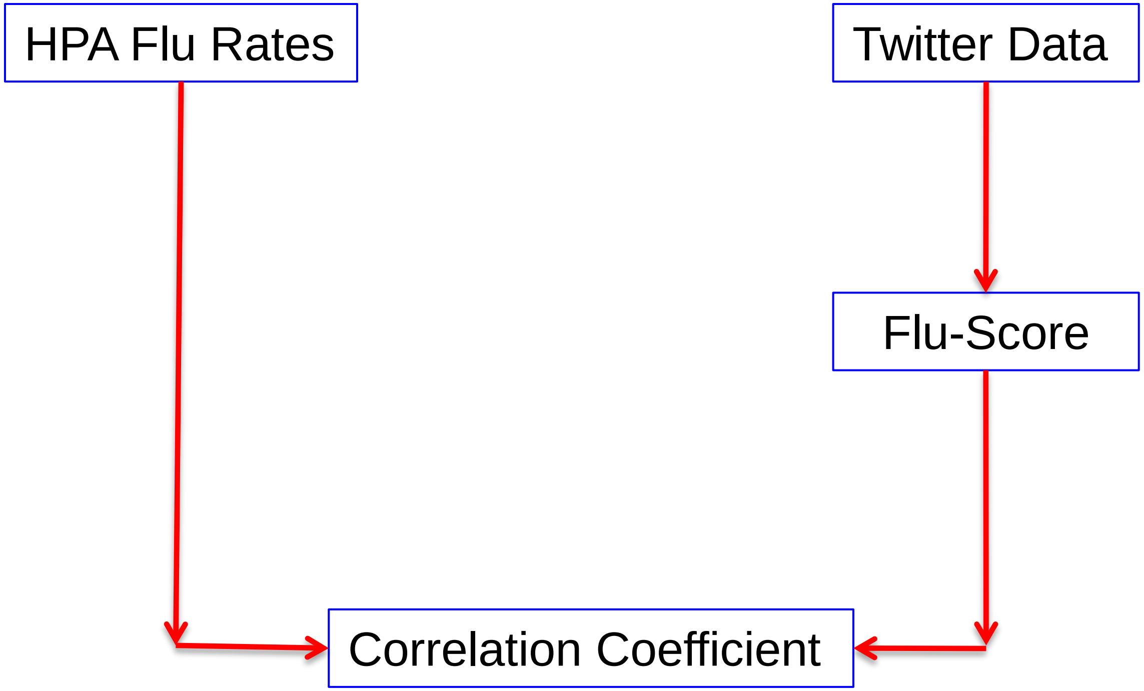


Fig. 1: Flu rates from the Health Protection Agency (HPA) for regions A-E (weeks 26-49, 2009). The original weekly HPA's flu rates have been expanded and smoothed in order to match with the daily data stream of Twitter (see section III-B).

Methodology



Computing Flu-scores

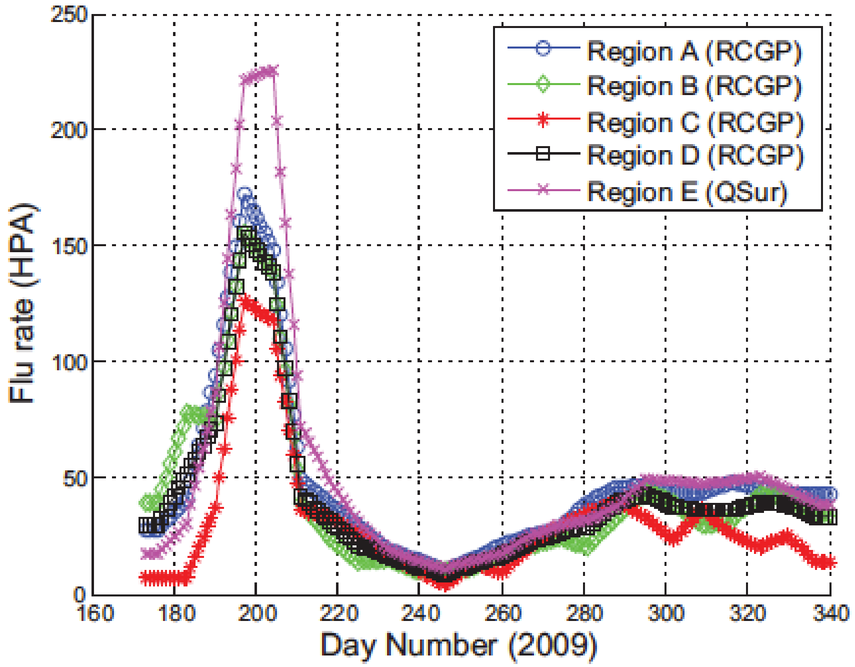
- The daily set of Tweets:

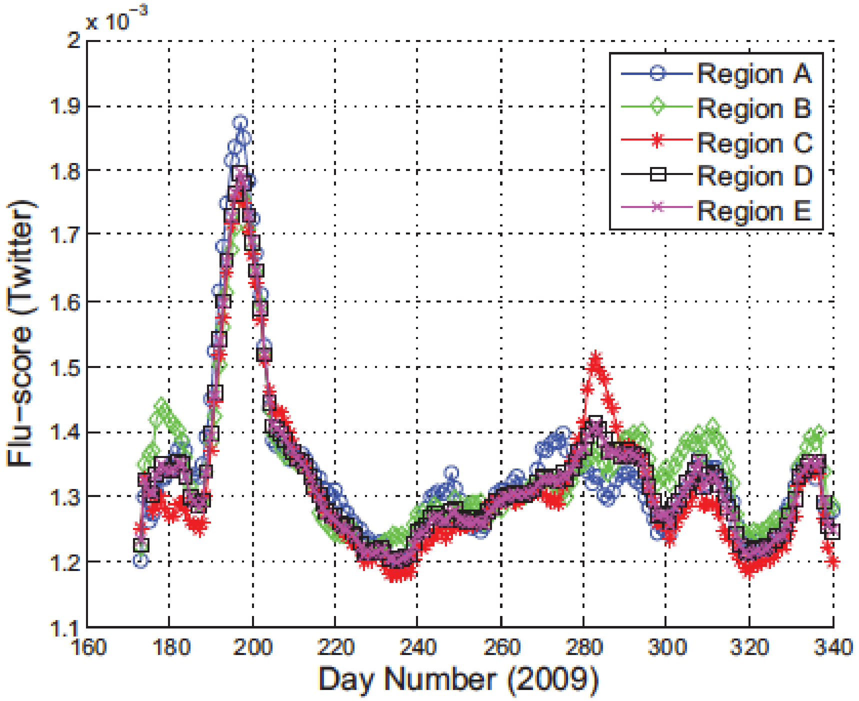
$$\mathcal{T} = \{t_j\}, \text{ where } j \in [1, n]$$

- Textual markers: expressing illness symptons, e.g. fever, temperature, sore throat, infection, headache
- A set of textual markers: $\mathcal{M} = \{m_i\} \quad i \in [1, k]$
  - Let $m_i(t_j)=1$ if m_i appears in tweet t_j, otherwise=0
- The flu-score of a tweet t_j: $s(t_j) = \frac{\sum_i m_i(t_j)}{k}$
- The flu-score of one day Twitter corpus

$$f(T,M) = \frac{\sum_j s(t_j)}{n} = \frac{\sum_j \sum_i m_i(t_j)}{k \bullet n}$$

Flu-rate vs Flu-score





Correlation Coefficient

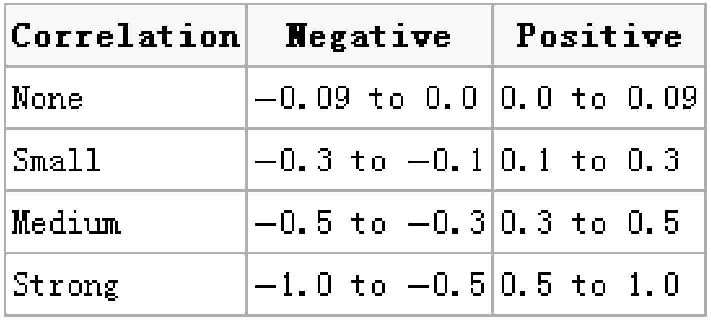
- A measure of the correlation (linear dependence) between two variables X and Y
- Giving a value between +1 and −1 inclusive
- Definition

$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

- For a sample:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}},$$

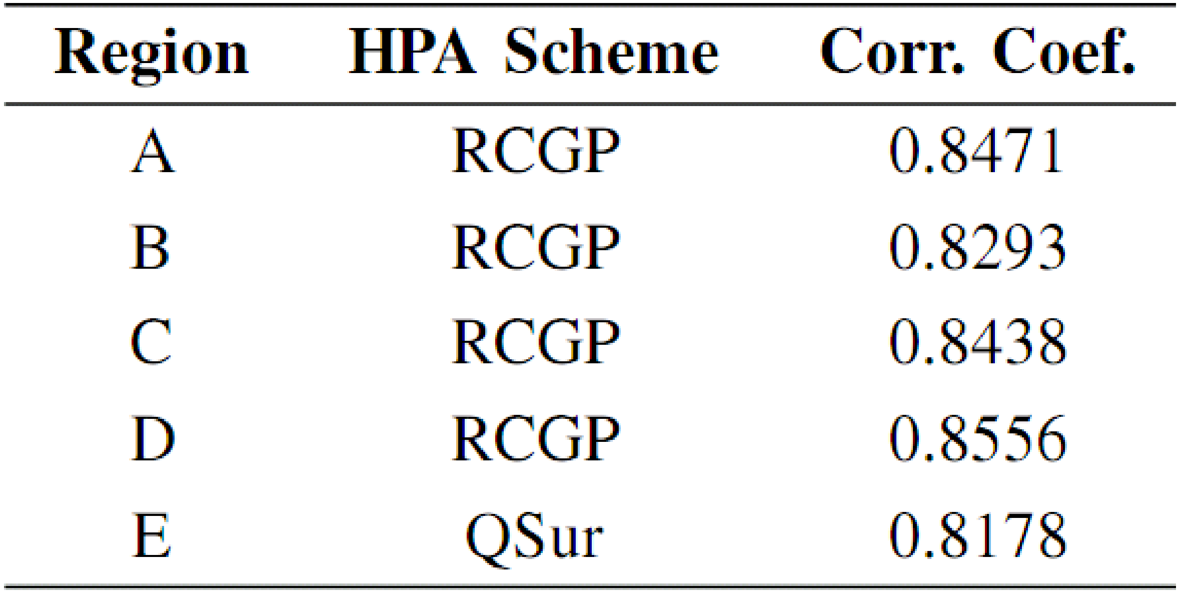
]



Correlation	Negative	Positive
None	−0.09 to 0.0	0.0 to 0.09
Small	−0.3 to −0.1	0.1 to 0.3
Medium	−0.5 to −0.3	0.3 to 0.5
Strong	−1.0 to −0.5	0.5 to 1.0

Correlations between Twitter Flu-scores and HPA Flu rates

- Strong correlation is observed!
- It indicates linear correlation between Twitter flu-scores and HPA flu rates, thus the flu-scores can be used to predict flu rates!



Region	HPA Scheme	Corr. Coef.
A	RCGP	0.8471
B	RCGP	0.8293
C	RCGP	0.8438
D	RCGP	0.8556
E	QSur	0.8178

Extensions

- Learning HPA’s flu rates from Twitter flu-scores
  - Linear regression is used to build a weighted Twitter flu-scores to model flu rates
- Automatic extraction of ILI textual markers
  - Creating candidate markers from:
    - Encyclopedic reference
    - Informal references
  - Forming the flu-subscores with time series.

Tracking Flu: Summary

- Tracking the flu outbreak in the UK using Twitter messages.
- High correlation between the flu-score and the HPA flu rates, greater than 95%.
- Advantages:
  - Less resource-demanding: only monitoring Twitter website, can be done automatically
  - More faster: can be done efficiently, while official reports need to delay 1 or two weeks
- Disadvantages
  - Still need sample data from official statistics for learning
  - Not everyone post their disease: could be not accurate
  - Not suitable for all sort of contagious diseases: some of them are not discussed in Twitter for privacy reason

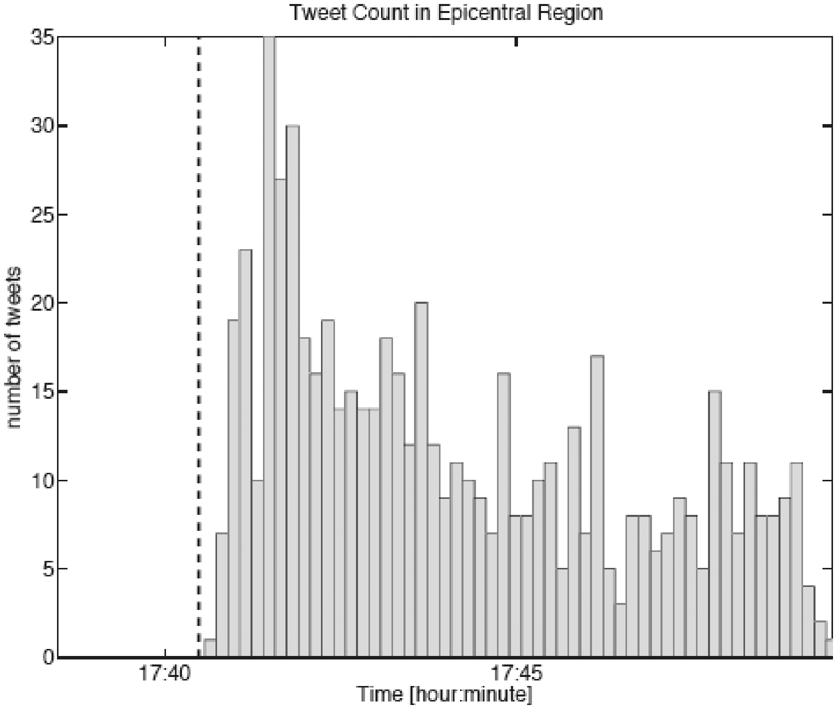
Tracking Earthquake Using Twitter [3,4]

[3] Paul Earle, Michelle Guy, Richard Buckmaster, Chris Ostrum, Scott Horvath, and Amy Vaughan, OMG Earthquake! Can Twitter Improve Earthquake Response? U.S. Geological Survey, 2010

[4] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web (WWW '10). ACM, New York, NY, USA, 851-860.

Part of our slides are from the authors of [4]

- Subsequent earthquakes generated volumes of earthquake-related tweets
- Access to firsthand accounts of earthquake shaking within seconds of an earthquake is intriguing



Tweet count following the 2009 MW 4.3 Morgan Hill, CA, earthquake. Tweets are binned in 10-second intervals, and the dashed line marks the origin time of the earthquake. After the earthquake, the tweet frequency quickly rose above the background level of less than one per hour to about 150 per minute.

Using Twitter for Earthquake Detection and Assessment

- Real-time nature of Microblogging
  - We can know what happens around other users in realtime
- Public tweets are stored in an openly searchable database
- Earthquake Reports
  - U.S. response time: 1.5-20 min
- Earthquake detection using Twitter
  - The typical delay for tweet transmission is 5 seconds
  - Earthquake could be detected in under a minute
- Twitter could be faster than Earthquake Wave!
  - An earthquake propagates at about 3–7 km/s (for 100km, about 20s)
  - Early alarm is possible?

Mapping the Felt Region

- Morgan Hill, 30 March 2009, MW 4.3

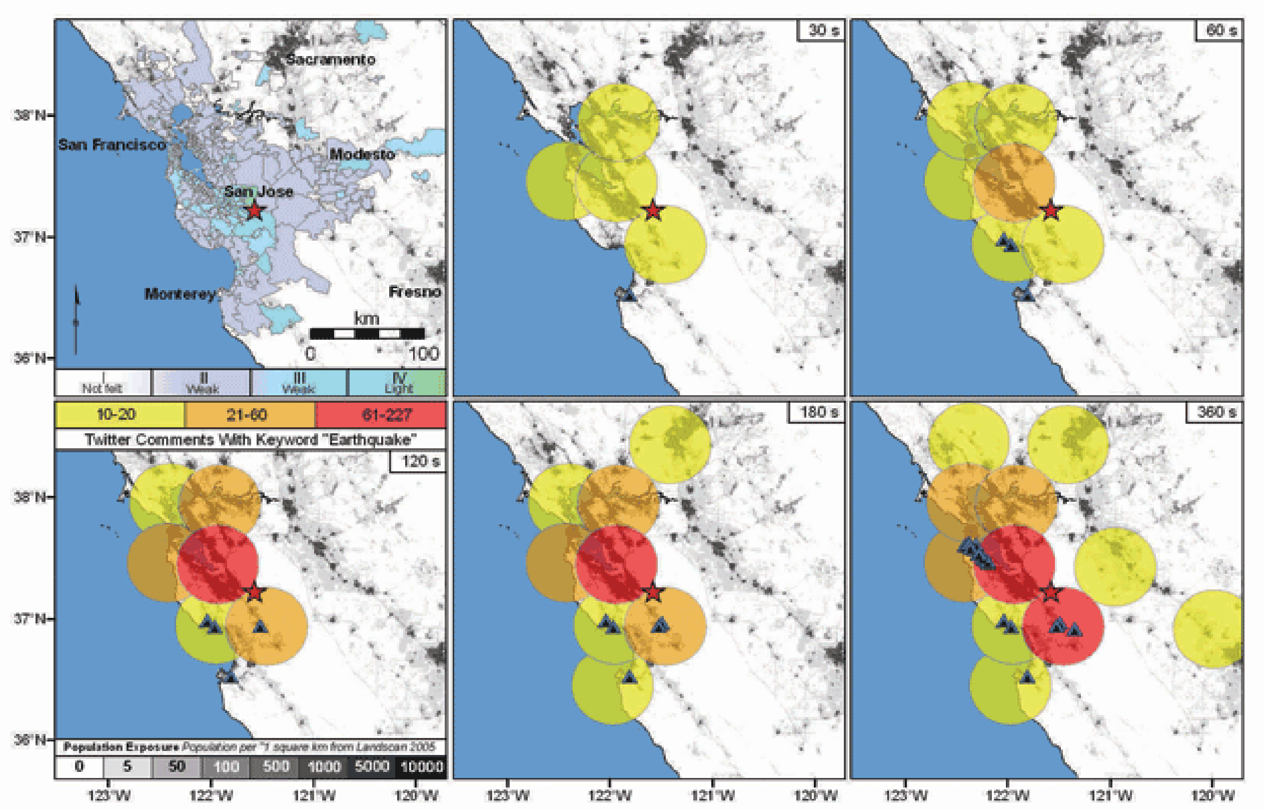
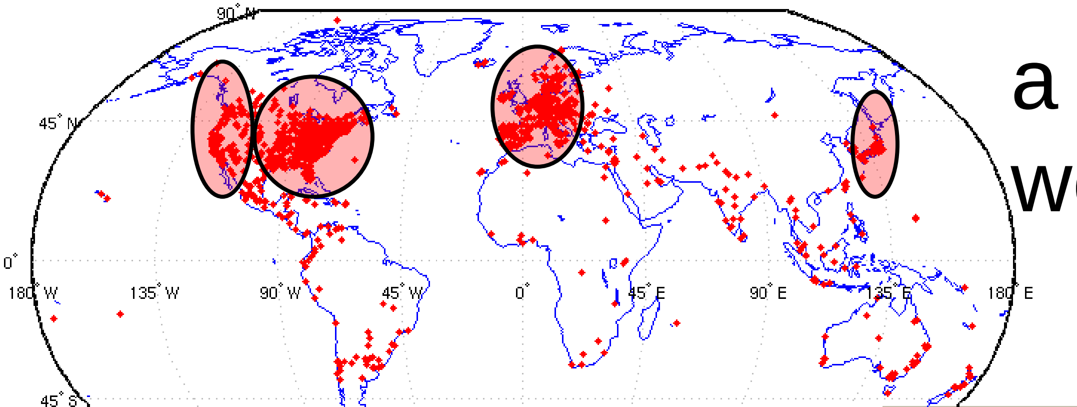


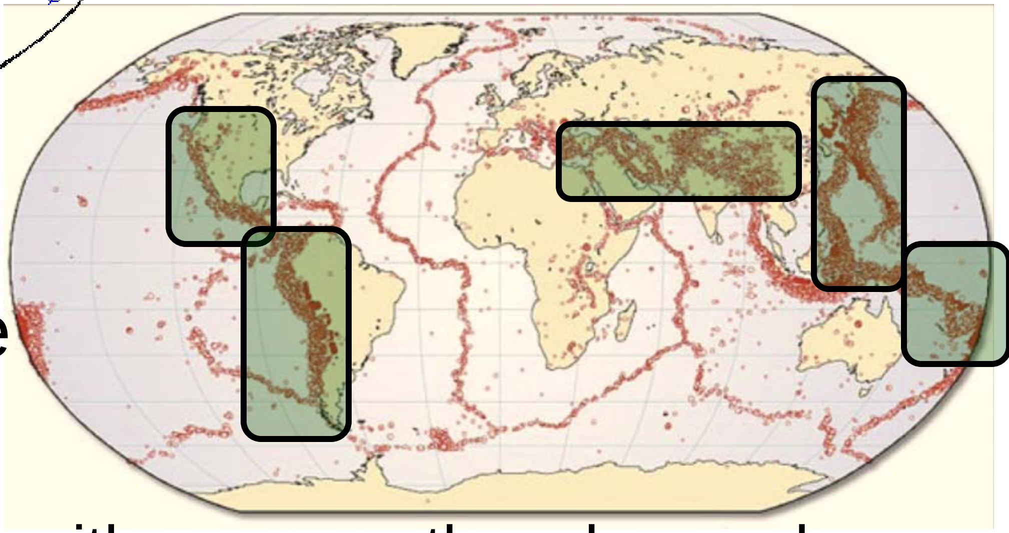
Figure 2. Comparison of the intensity map produced by the USGS DYFI system (upper left) and the geocoded tweet counts for the 30 March 2009 MW 4.3 Morgan Hill earthquake. The extent of the geographic tweet search is indicated by the size of the circles, which are color coded by the number of tweets. The population is shown in the background as gray scale and the tweets with exact latitude and longitude geo-references are shown as black triangles with blue outlines. The different panels show the integrated tweet count at discrete times after the earthquake as indicated in the upper right corner of the maps.

Slides from Takehi Sakaki et. al.

Twitter and Earthquakes in Japan



a map of Twitter user world wide



a map of earthquake occurrences world wide

The intersection is regions with many earthquakes and large twitter users in Japan.

Other regions:

Indonesia, Turkey, Iran, Italy, and Pacific coastal US cities

Event Detection Using Twitter

- Do semantic analysis on Tweet
  - To obtain tweets on the target event precisely

- Regard Twitter user as a sensor
  - To detect the target event
  - To estimate location of the target

Semantic Analysis on Tweet

- Search tweets including keywords related to a target event
  - Example: In the case of earthquakes
    - "shaking", "earthquake"

- Classify tweets into a positive class or a negative class
  - Example:
    - "Earthquake right now!!" ---positive
    - "Someone is shaking hands with my boss" --- negative
  - Create a classifier

Semantic Analysis on Tweet

- Create classifier for tweets
  - use Support Vector Machine(SVM) - a machine learning algorithm

- Features (Example: I am in Japan, earthquake right now!)
  - Statistical features (7 words, the 5th word)

    the number of words in a tweet message and the position of the query within a tweet

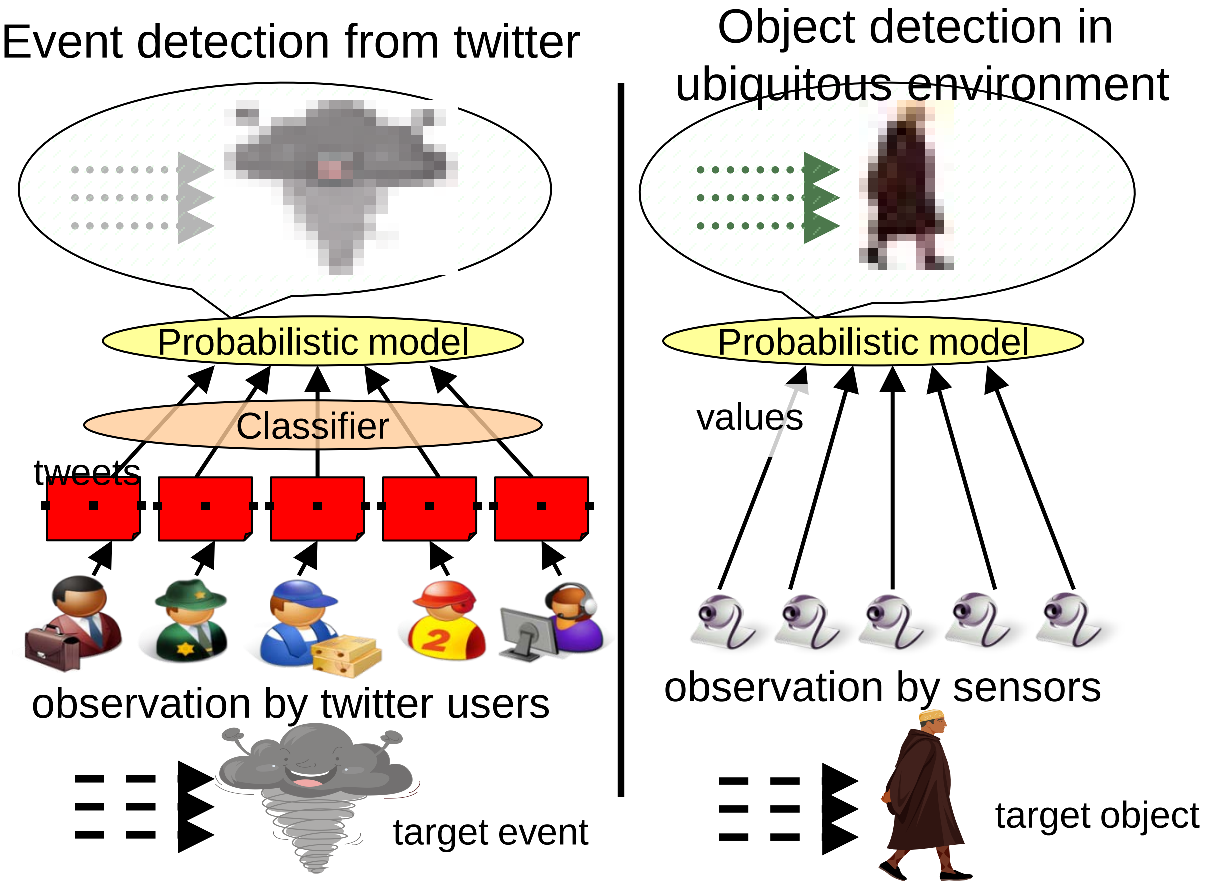
  - Keyword features (I, am, in, Japan, earthquake, right, now)

    the words in a tweet

  - Word context features (Japan, right)

    the words before and after the query word

Tweet as a Sensory Value



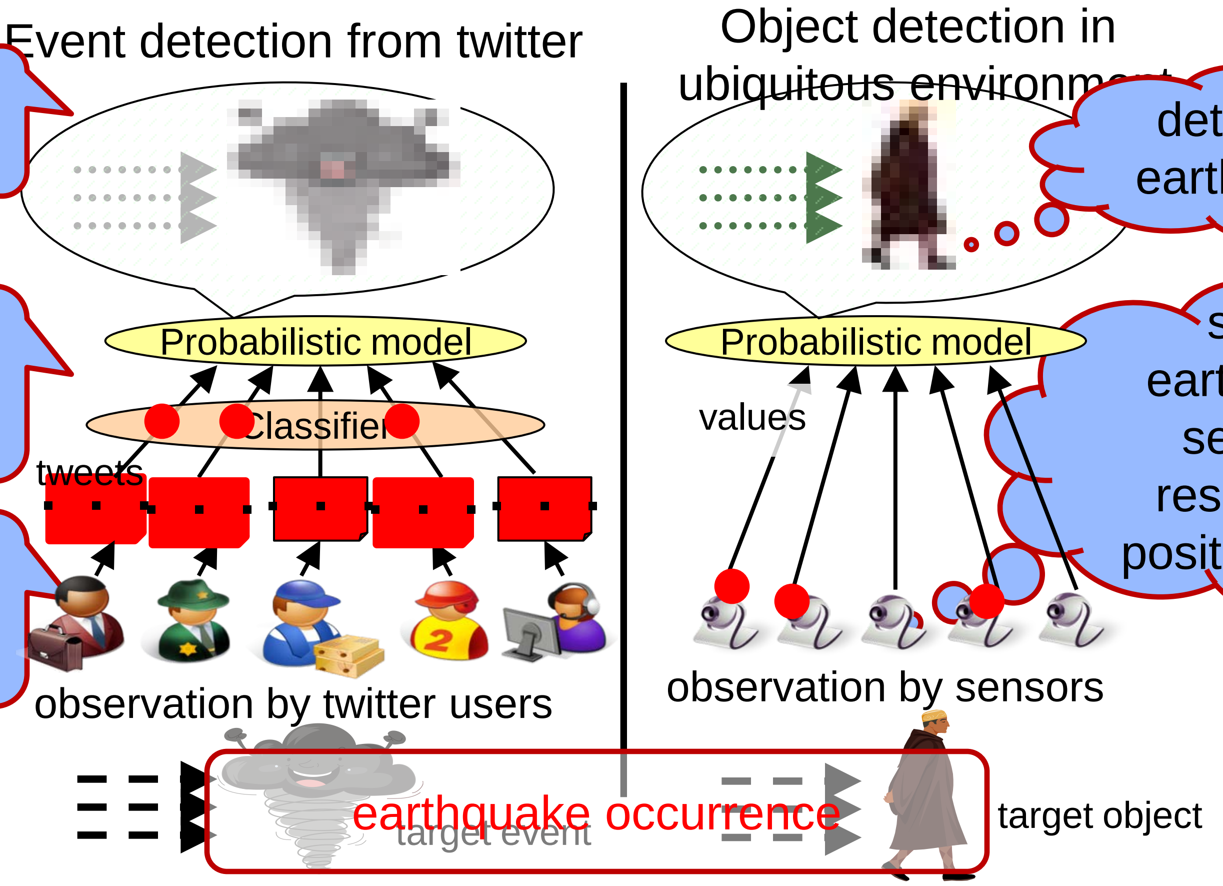
the correspondence between tweets processing and sensory data detection

Tweet as a Sensory Value

detect an earthquake

search and classify them into positive class

some users posts “earthquake right now!!”



detect an earthquake

some earthquake sensors responses positive value

We can apply methods for sensory data detection to tweets processing

Tweet as a Sensory Value

- We make two assumptions to apply methods for observation by sensors

- Assumption 1: Each Twitter user is regarded as a sensor
  - a tweet →a sensor reading
  - a sensor detects a target event and makes a report probabilistically
  - Example:
    - make a tweet about an earthquake occurrence
    - “earthquake sensor” return a positive value

- Assumption 2: Each tweet is associated with a time and location
  - a time : post time
  - location : GPS data or location information in user’s profile

Processing time information and location information, we can detect target events and estimate location of target events

Earthquake Location Estimation

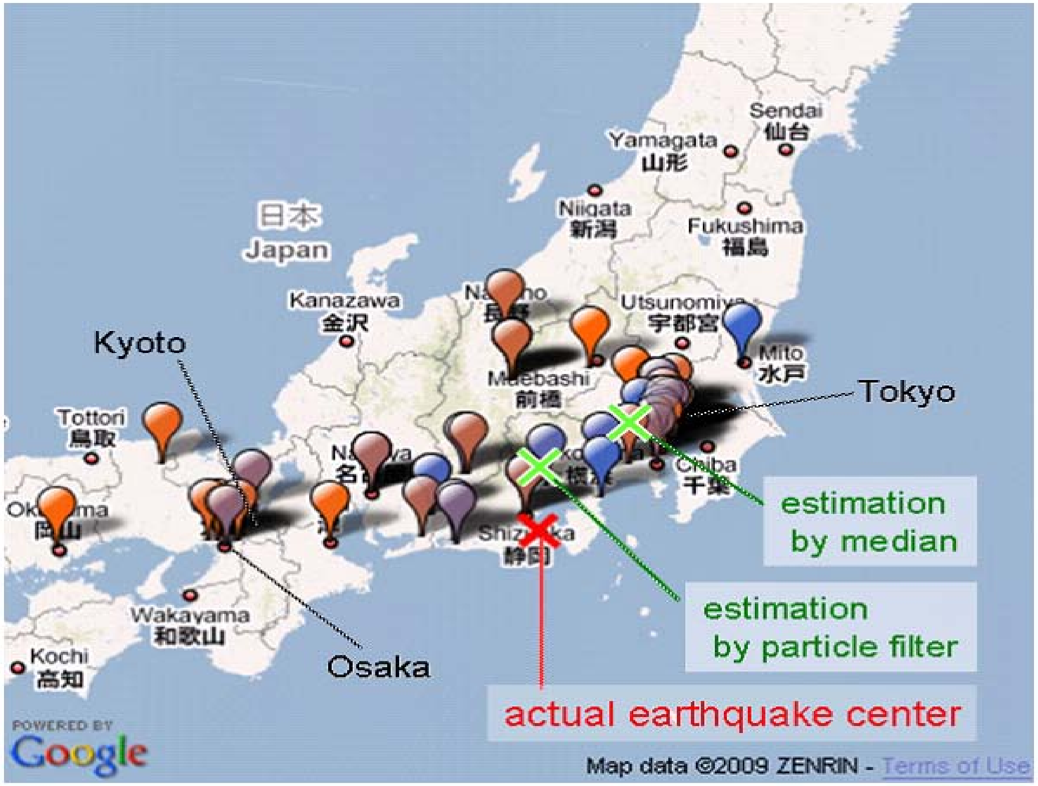
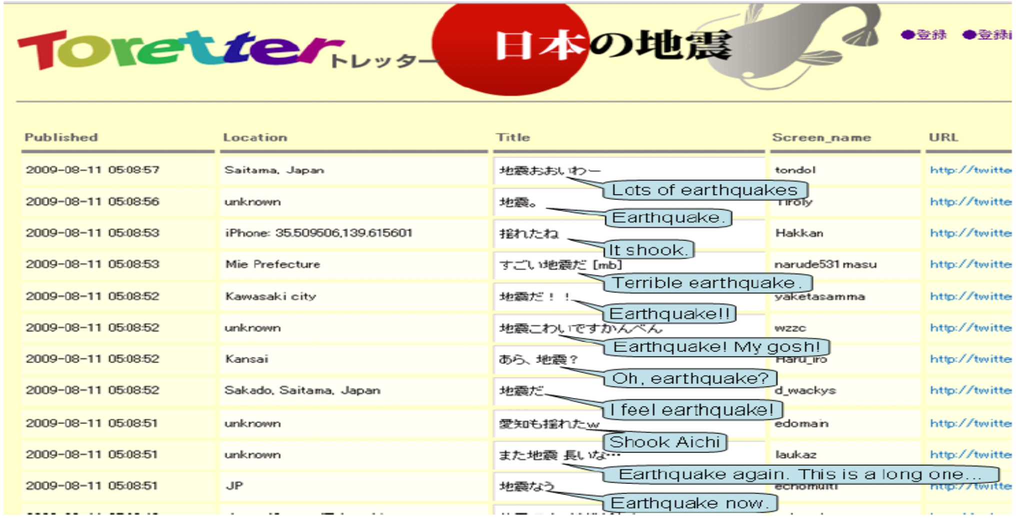


Figure 9: Earthquake location estimation based on tweets. Balloons show the tweets on the earthquake. The cross shows the earthquake center. Red represents early tweets; blue represents later tweets.

Earthquake Reporting System

- Toretter (http://toretter.com)
  - Earthquake reporting system using the event detection algorithm
  - All users can see the detection of past earthquakes
  - Registered users can receive e-mails of earthquake detection reports



Published	Location	Title	Screen_name	URL
2009-08-11 05:08:57	Saitama, Japan	地震おおいわー Lots of earthquakes	tondol	http://twitte
2009-08-11 05:08:56	unknown	地震。 Earthquake.	Tiroly	http://twitte
2009-08-11 05:08:53	iPhone: 35.509506,139.615601	揺れたね It shook.	Hakkan	http://twitte
2009-08-11 05:08:53	Mie Prefecture	すごい地震だ [mb] Terrible earthquake.	narude531masu	http://twitte
2009-08-11 05:08:52	Kawasaki city	地震だ！！ Earthquake!!	yaketasamma	http://twitte
2009-08-11 05:08:52	unknown	地震こわいですかんべん Earthquake! My gosh!	wzzc	http://twitte
2009-08-11 05:08:52	Kansai	あら、地震？ Oh, earthquake?	Haru_iro	http://twitte
2009-08-11 05:08:52	Sakado, Saitama, Japan	地震だ I feel earthquake!	d_wackys	http://twitte
2009-08-11 05:08:51	unknown	愛知も揺れたw Shook Aichi	edomain	http://twitte
2009-08-11 05:08:51	unknown	また地震 長いな… Earthquake again. This is a long one…	laukaz	http://twitte
2009-08-11 05:08:51	JP	地震なう Earthquake now.	echomulti	http://twitte

Earthquake Reporting System

- Effectiveness of alerts of this system
  - Alert E-mails urges users to prepare for the earthquake if they are received by a user shortly before the earthquake actually arrives.

- Is it possible to receive the e-mail before the earthquake actually arrives?
  - An earthquake is transmitted through the earth's crust at about 3~7 km/s.
  - a person has about 20~30 sec before its arrival at a point that is 100 km distant from an actual center

Results of Earthquake Detection

- In all cases, we sent E-mails before announces of JMA
- In the earliest cases, we can sent E-mails in 19 sec.

Date	Magnitude	Location	Time	E-mail sent time	time gap [sec]	# tweets within 10 minutes	Announce of JMA
Aug. 18	4.5	Tochigi	6:58:55	7:00:30	95	35	7:08
Aug. 18	3.1	Suruga-wan	19:22:48	19:23:14	26	17	19:28
Aug. 21	4.1	Chiba	8:51:16	8:51:35	19	52	8:56
Aug. 25	4.3	Uraga-oki	2:22:49	2:23:21	31	23	2:27
Aug.25	3.5	Fukushima	2:21:15	22:22:29	73	13	22:26
Aug. 27	3.9	Wakayama	17:47:30	17:48:11	41	16	1:7:53
Aug. 27	2.8	Suruga-wan	20:26:23	20:26:45	22	14	20:31
Ag. 31	4.5	Fukushima	00:45:54	00:46:24	30	32	00:51
Sep. 2	3.3	Suruga-wan	13:04:45	13:05:04	19	18	13:10
Sep. 2	3.6	Bungo-suido	17:37:53	17:38:27	34	3	17:43

Discussion

- Advantages
  - No need of dedicate devices
  - Provide a fast detection and assessment of earthquake
  - Possibility of early alarm
- Limitations of earthquake detection with Twitter
  - Need enough Twitter samples
    - Depend on the population distribution of Twitter
    - If the center of a target event is in an oceanic area, it's more difficult to locate it
    - The number of tweets maybe not as large as we had anticipated.
  - Could be not accurate
    - Interfered by retweets (not in the earthquake area)
    - Incorrect tweet geolocations
  - Could be unstable
    - The network service is not reliable when earthquake happens
    - Vulnerable to hacker attacks
  - Still not fast enough

Other Possible Applications

- Social life
  - Detect the hot news in the world
- Economy
  - Detect the trend of stocks
- Politics
  - Predict and evaluate president selection and other political events
- Science
  - Discover patterns of social interactions and influences
- …

Summary

- Applications
  - Decentralized search
  - Epidemics
    - SIRS Model
    - Flu detection
  - Tracking Earthquake Using Twitter