Information Diffusion in Social Networks

Research Promotion Workshop
BESU, Shibpur
15th March 2013

Amitabha Bagchi
Computer Science and Engineering
IIT Delhi

Online Social Networks

- OSNs like Facebook and Twitter are ubiquitous.
- In fact some of you are probably updating your Facebook status even as I speak.
- "Stuck in boring talk about research, think I'll take a nap....LOL"

Researchers from various disciplines are waking up to the possibilities.

Research aspects of OSNs

- Sociologists have studied human social networks from the dawn of their discipline.
- Physicists are interested in social networks as a complex system of interacting agents
- Mathematicians see stochastic processes.
- Economists apply game theory

Computer Scientists built these systems. And we are building the systems that can analyze the data these systems generate.

Information diffusion on OSNs

Question: How do particular topics or pieces of content become popular on OSNs?

The answer to this question is tremendously important to a variety of stakeholders: commerce, political scientists, sociologists etc

Two aspects: Macro and Micro

Micro: What are individual users doing?

Macro: What are the large-scale phenomena that are observed in this system?

Synthesis: Can we deduce the nature of the large-scale phenomena from a knowledge of what individual users are doing?

Example: The SIR model

Given a graph **G** and a special vertex **v** that has a certain message (rumor).

- Each node is in one of three states: Susceptible, Infected, Removed. Initial **v** is Infected and everyone else is Susceptible.
- At each time step an edge **(u,v)** is chosen at random and if **u** is infected it sends the message to **v.**
- If **v** is **S**, it becomes **I**. If it is **I** it becomes **R**.
- If **v** is **R** then **u** becomes **R.**

SIR: The Macro question

Clearly, as long as there are infected nodes the process continues.

Question: Will all the nodes have been infected for at least some time before the process ends?

Ans: (Probably) depends on the topology. For a complete graph the answer is no (Sudbury, J. Appl. Prob., 1985).

The way of Physics

Observe the macro and theorize about the micro to better understand the universe.

The way of Engineering

Use the observation of the micro and the theory of the micro to build better systems and make more money...

...thereby helping pay for Physics research

Outline

- Refine the micro question.
- Define a stochastic model of the micro.
- Simulate and observe the behaviour of the macro.
- Compare with data.

Refining the question

The Attribution problem: Why do users do what they do?

- Did you share that photo because you like what's in it or because you are a big fan of the person who posted it?
- You just heard on TV that Sehwag has been cut from the Indian team. Do you want to share your opinion on Twitter?
- Everyone is talking about Kolaveri. Do you want to check it out?

Building the model

The model comes from (possible) answers to the questions.

- People are influenced by what their friends are talking about. (*Endogenous*).
- People monitor broadcast media also and often respond to it on OSNs. (*Exogenous*).
- People respond to themes that are getting popular on OSNs. (*Somewhere in between*).

The Model I

- Users form a network that is an undirected small-world.
- Each user “tweets” from time to time. A “tweet” is an event in time that has a “topic” associated with it.
- The users options of topics at time t are from a set of topics that have been seen until time t.
- The user differentiates between “global” topics and “local” topics.

The Model II

- There is a “global list” in which “global tweets” arrive with frequency λ_1 (distributed as a Poisson point process). Each of these brings a new topic.
- Each user has a “local list” into which tweets are written with frequency λ_2 (distributed as a Poisson point process).

The topic of a user’s tweet is chosen randomly out of the topics in the global list and the local lists of its neighbours in the network.

The Model III

- Each global tweet has a weight A on arrival in the global list.
- This weight decreases exponentially with time with parameter α i.e. $Ae^{-\alpha t}$ at time t if the topic arrived at time 0.
- When a user tweets then that tweet is placed in its local list with weight B.
- This weight decreases exponentially with time with parameter β i.e. $Be^{-\beta t}$ at time t if the tweet arrived at time 0.

The Model IV

A new tweet has two kinds of candidates it can copy its topic from:

- Global tweets.
- Local tweets from one of its neighbours' lists.

A new tweet has the same topic as a candidate tweet with probability proportional to the candidate's weight.

A reality check

- The total weight seen by any node is finite with probability 1.
- Additionally since this is an ergodic Markov process there is a stationary distribution, hence the total weight converges to a constant C(v) for node v.

$$E[C] = \lambda_1 A/\alpha + k\lambda_2 B/\beta,$$

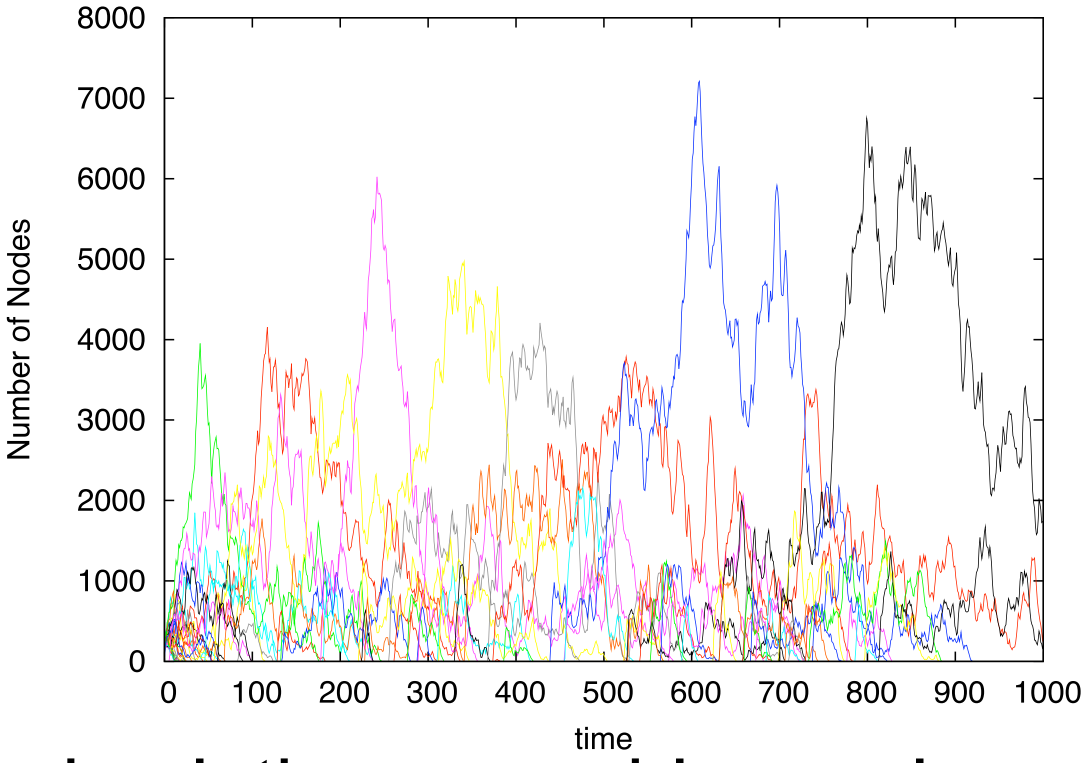
Where k is the number of neighbors of v.

Three parameter regimes

Varying the parameters gives us three kinds of behaviours.

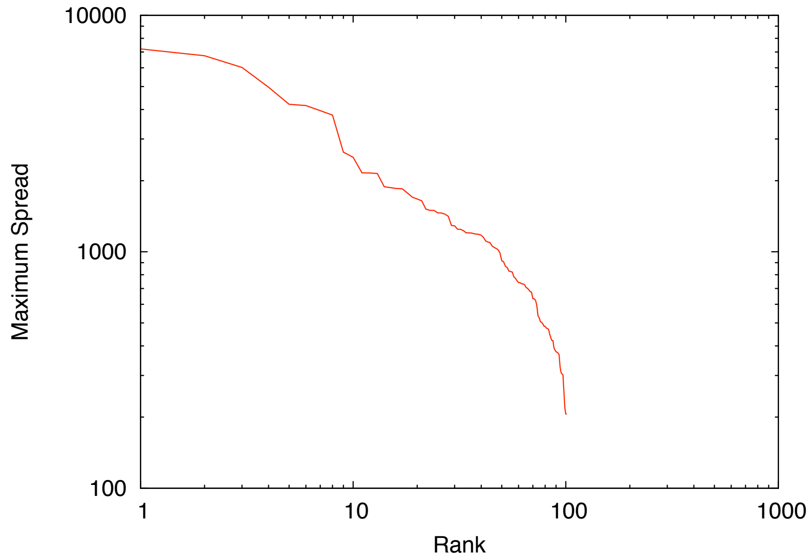
- Sub-viral regime: No topic dominates. Well-described by mean-field approximation.
- Super-viral regime: Each new topic goes viral and dies quickly
- Viral regime

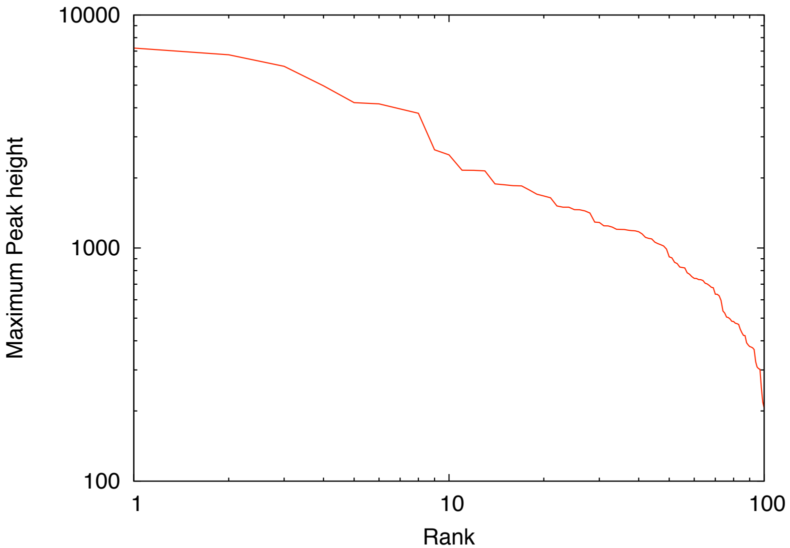
Evolution in the viral regime



The simulation resembles real-world topic evolution.

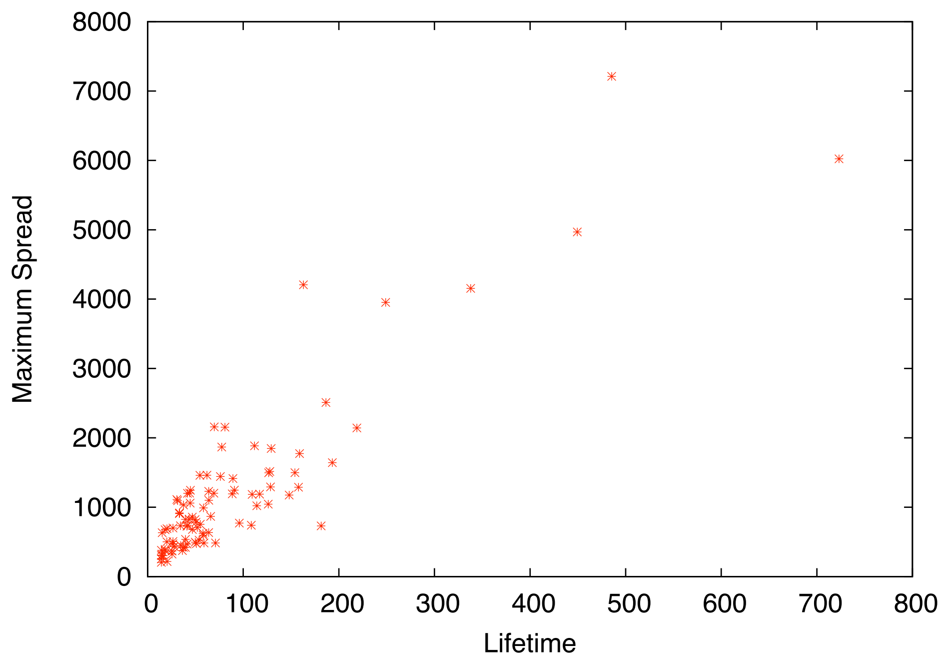
Viral regime characteristics





Power law-like distributions are seen for macro properties like peak volume, spread and lifetime.

Live longer, go further



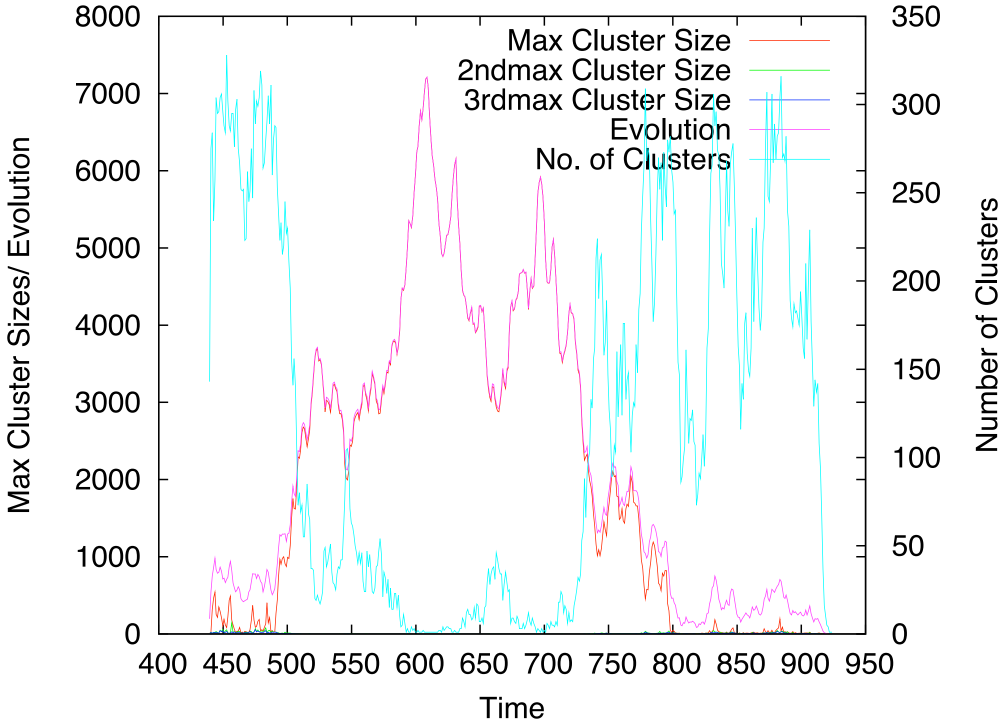
Longer lived topics spread further. (Or is it the other way around?)

Studying topology empirically

We define topic based graphs for each topic

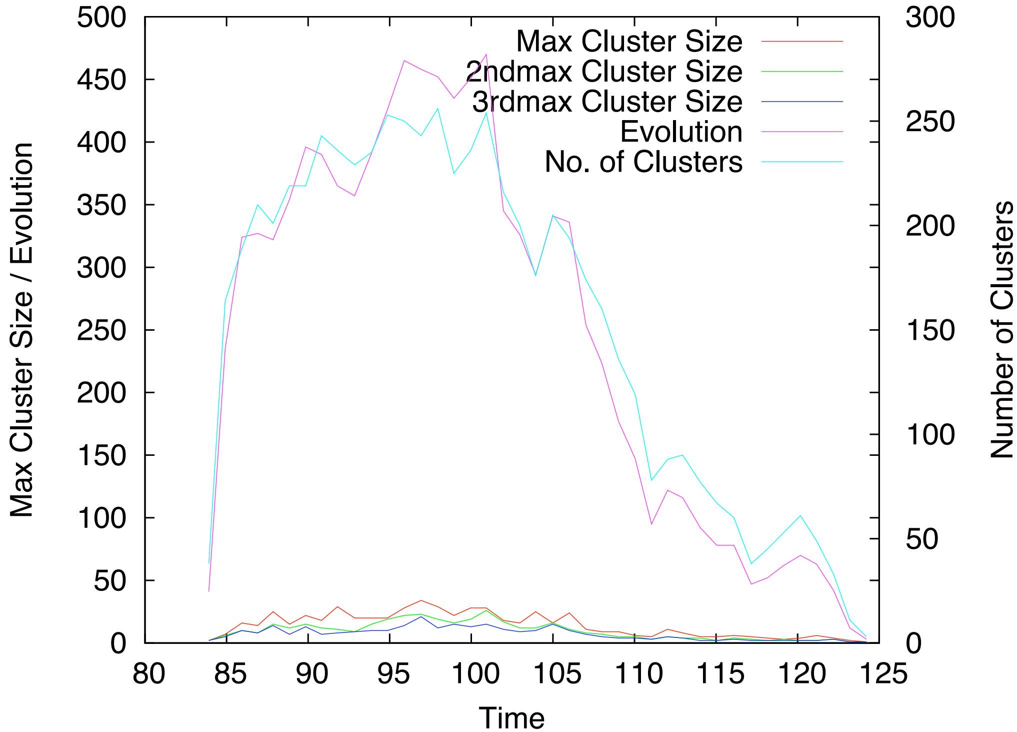
- **Lifetime graph:** The subgraph induced by all users who have ever tweeted on the topic.
- **Evolving graphs:** The sequence of graphs induced by the users who tweet on the topic on a given day.
- **Cumulative evolving graph:** There is an edge from **u** to **v** if **u follows v and u tweets on the topic a day after** tweets on day **t** and

Topological study: Viral topics



For a viral topic clusters merge into one as it rises in popularity. (Evolving graph)

Topological study: Non-viral topics

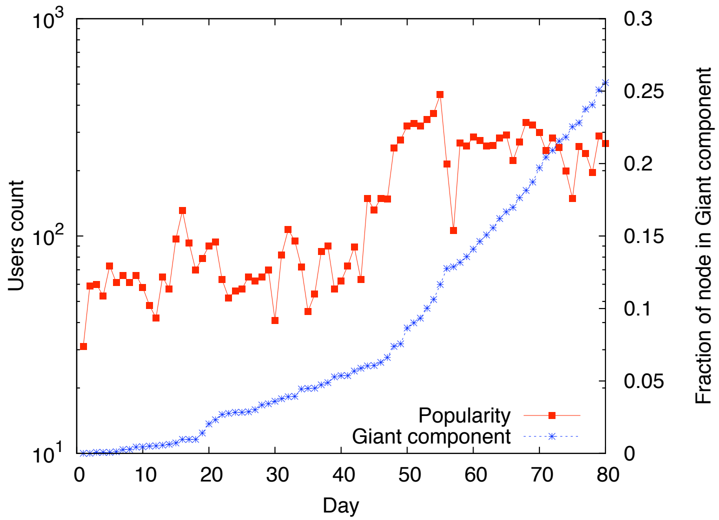


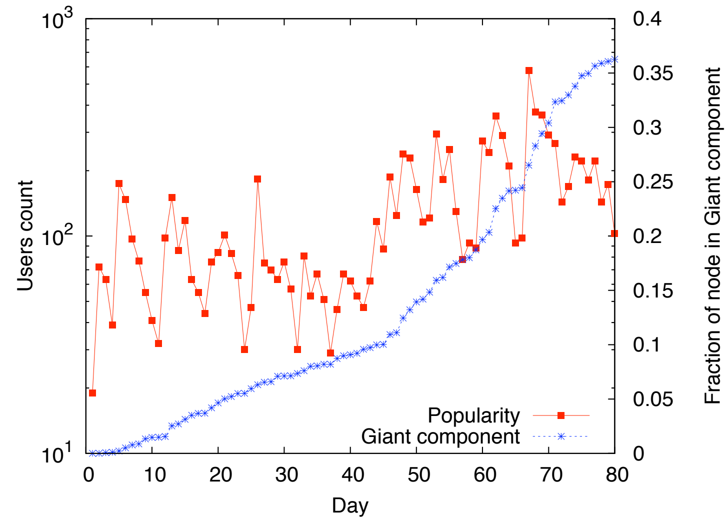
Non-viral topics see many small clusters.
(Evolving graph)

Empirical cross-verification: Setup

- We used a data set containing approx 200 million tweets from 9 million users crawled from Twitter in 2009.
- We augmented the data set by crawling follower-following relationships and geolocating the users where possible.
- Further we used NLP tools to tag tweets with topics (since hashtags were very sparse).

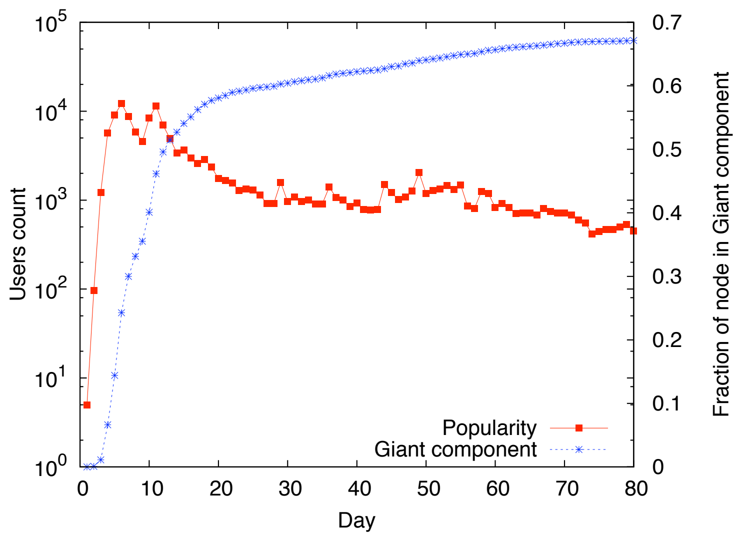
Large cluster formation: Empirical

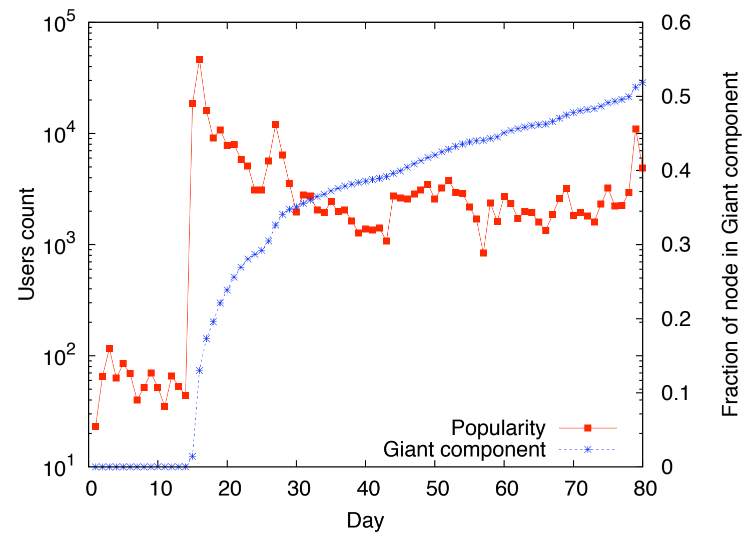




For **non-viral** topics, the largest component of the cumulative evolving graph contains a small fraction of all nodes

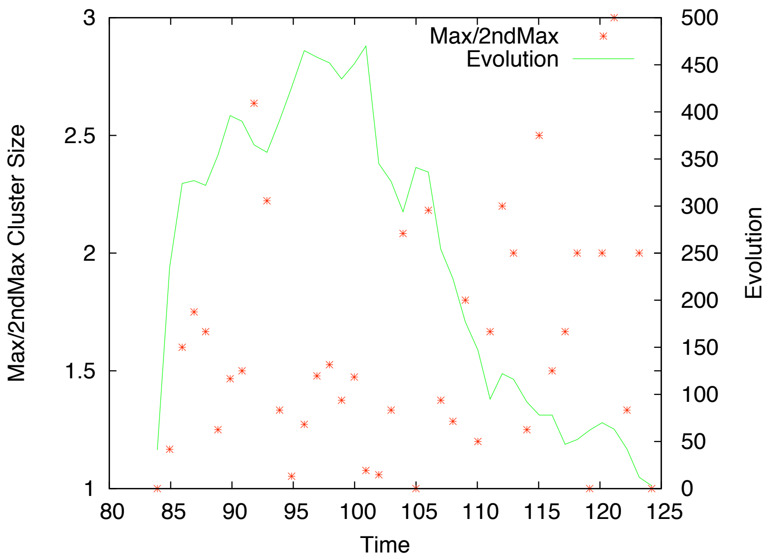
Large clusters in viral topics

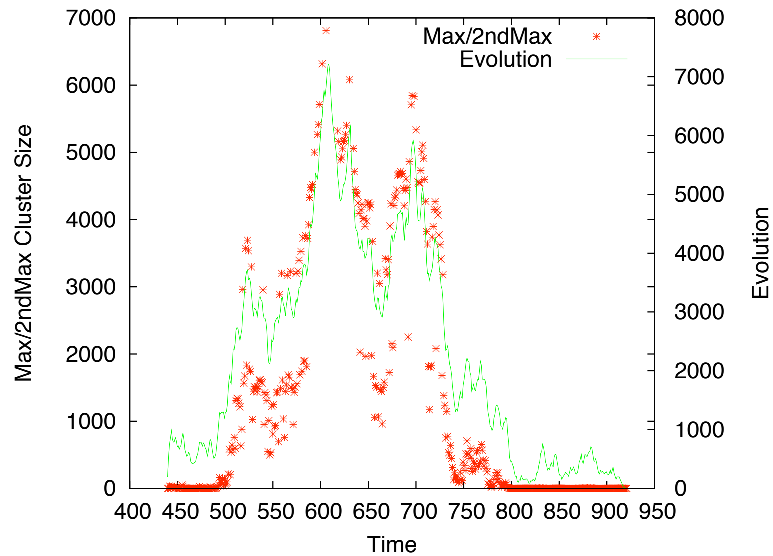




In viral topics the largest component takes up a significant fraction of the graph, growing in size as the topic rises in popularity.

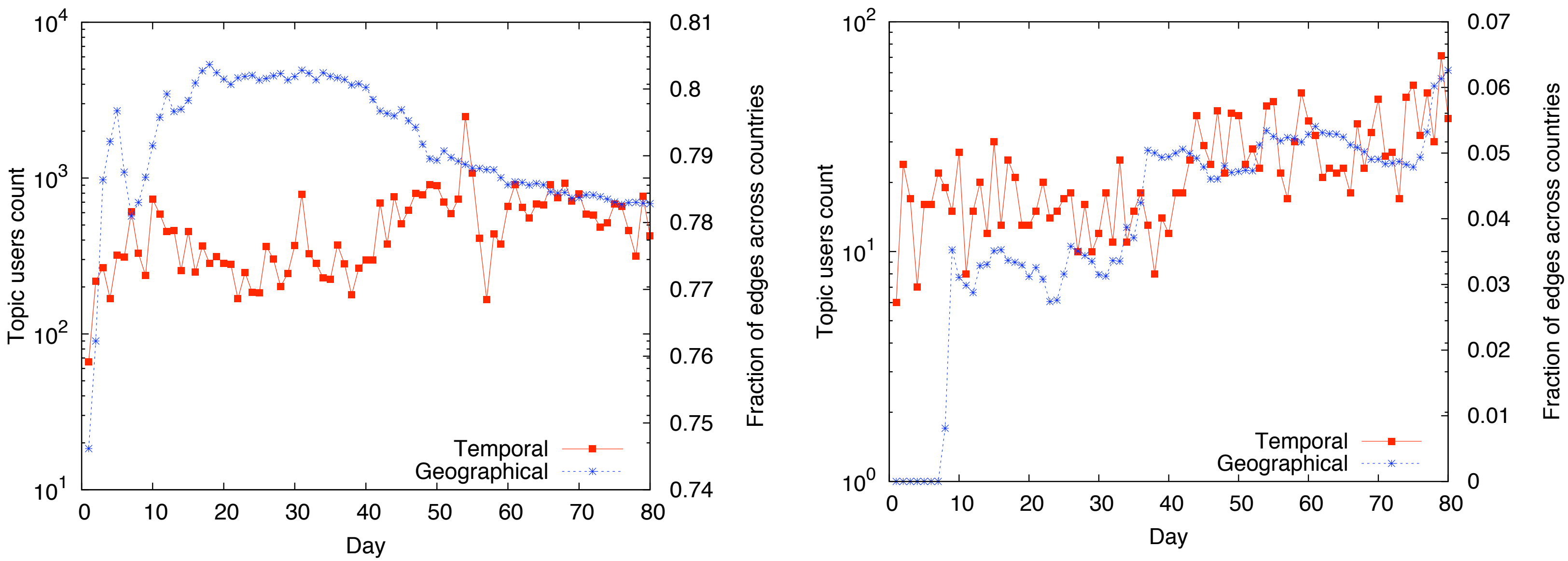
Cluster merging in the model





The ratio of the largest to the second largest component in the evolving graph tells a story.

The real data also has geography



Viral topics cross regional/national boundaries in the cumulative evolving graph.

That was the trailer…

- Ruhela et. al. Towards the use of Online Social Networks for Efficient Internet Content Distribution, in Proc ANTS 2011.
- Ardon et. al. Spatio-Temporal Analysis of Topic Popularity in Twitter, arXiv:1111.2904v1 [cs.SI].
- Rajyalakshmi et. al. Topic Diffusion and Emergence of Virality in Social Networks, arxiv: 1202.2215v1 [cs.SI].

www.cse.iitd.ernet.in/~bagchi

The emerging science of big data

- Huge amounts of data being generated from all kinds of sources.
- “Smart cities”, Genome sequencing, telescopes, networked systme.
- A growing awareness that the science of data is the new frontier of technology

Think of it as IT’s steam engine moment. It’s turn to shine as a force in human affairs.

Challenges

- Modelling
  - Domain knowledge required
  - But understanding of what data can reveal also required.
- Data analytics
  - Algorithms
  - Data structures
  - Databases
  - System Architecture

The research horizon..

- …is unlimited.
- Only CS fundamentals will matter.
- Everything else will become obsolete before the exam papers are returned.

Thanks for listening