

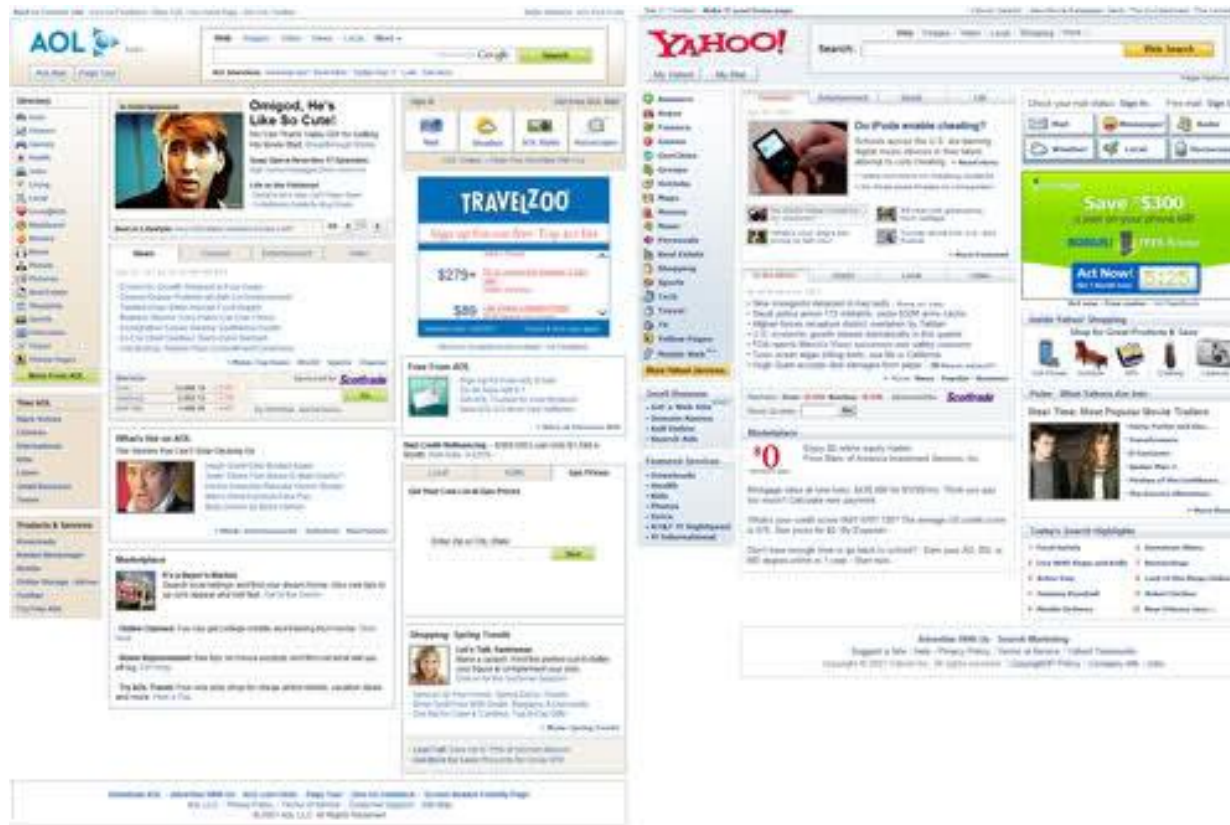
On Word-of-Mouth Based Discovery of the Web

Tiago Rodrigues (UFMG), Fabrício Benevenuto (UFOP), Meeyoung Cha (KAIST), Krishna P. Gummadi (MPI-SWS), Virgílio Almeida (UFMG)



Information Discovery on the Web

- ▶ Browsing
 - ▶ Aggregators of content and links





Information Discovery on the Web

▶ Searching





Information Discovery on the Web

- ▶ Social Media
 - ▶ Word-of-Mouth: extremely popular and globally reaching



Social Media and Information Discovery

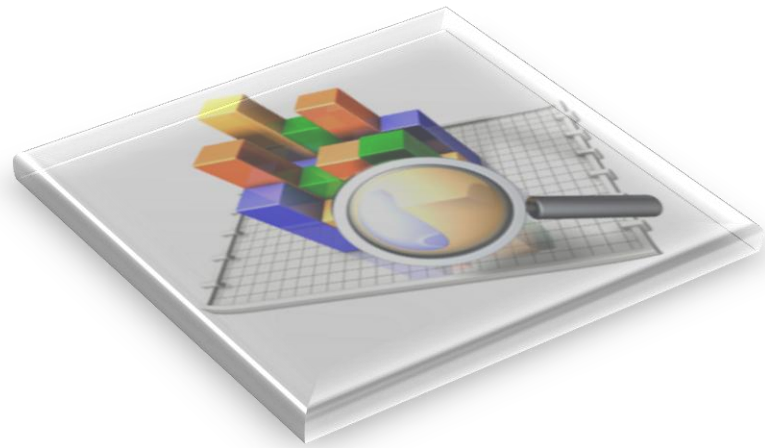
- ▶ **State of the Media: The Social Media Report**
[<http://blog.nielsen.com/nielsenwire/social/>]
 - ▶ Americans spent 23% of their online time on social media
 - ▶ Americans spent more time on Facebook than they do on any other website

- ▶ **OSN sites are a major driver of traffic to many web sites**
 - ▶ Nearly 23 million web links are shared every day on Twitter
Kwak et. al [What is Twitter, a Social Network or a News Media?, WWW, 2010]

Our Goal

- ▶ Large scale analysis of Word-of-Mouth web based content discovery on Twitter
- ▶ Interesting Research Questions
 - ▶ What types of content are discovered by Word-of-Mouth?
 - ▶ What are the structures of Word-of-Mouth propagation trees?
 - ▶ How geographically distributed are the propagation trees?

Methodology





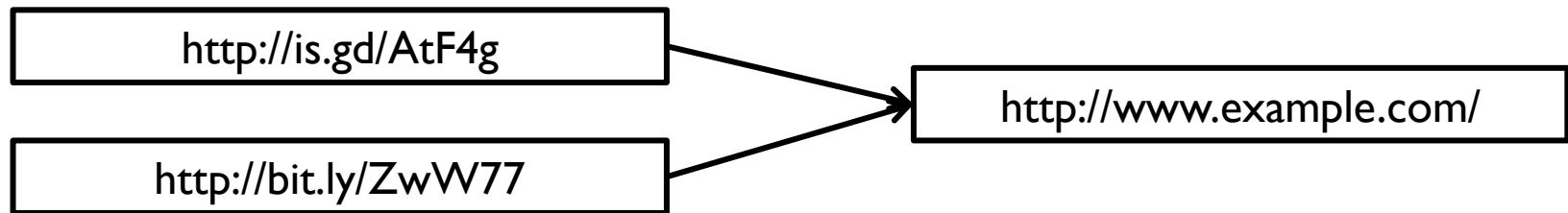
Why URLs on Twitter?

- ▶ Online social network and microblogging service
 - ▶ Users follow others to receive their 140 characters tweets
- ▶ Ideal medium to study Word-of-Mouth
 - ▶ Centered around the idea of spreading information
 - ▶ Additional mechanisms like retweet
 - ▶ URL shortening services
- ▶ 208M URLs were shared on Twitter from 2006 to 2009
 - ▶ Clean piece of information

Data Collection

▶ Samples

- ▶ Cha et. al [Measuring User Influence in Twitter: The Million Follower Fallacy, ICWSM, 2010]
- ▶ Compare URLs from the same period
- ▶ Performance issues on data management

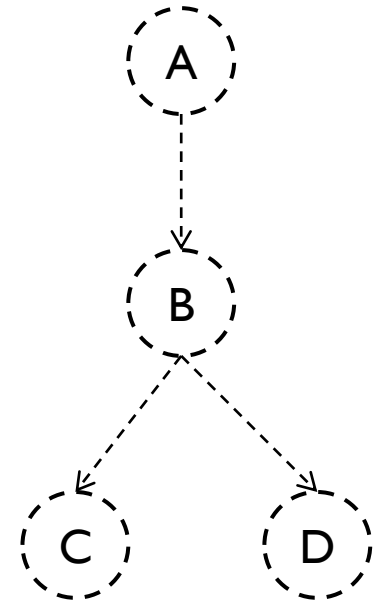


	Period	Distinct URLs	Tweets	Retweets	Users
Dataset 1	Jan 1, 2009 – Jan 7, 2009	1,239,445	6,028,030	295,665	995,311
Dataset 2	Apr 1, 2009 – Apr 7, 2009	4,628,095	17,381,969	1,178,244	2,040,932

Modeling Information Cascades

- ▶ Hierarchical tree model

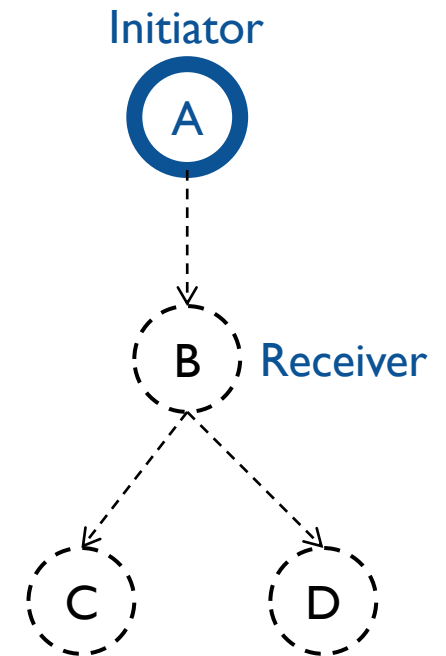
T	User	Tweet content



Modeling Information Cascades

- ▶ Hierarchical tree model

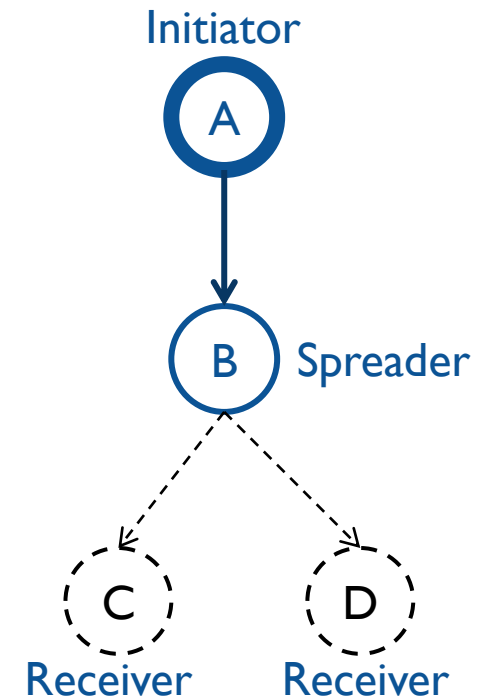
T	User	Tweet content
1	A	Check this: http://www.example.com/



Modeling Information Cascades

► Hierarchical tree model

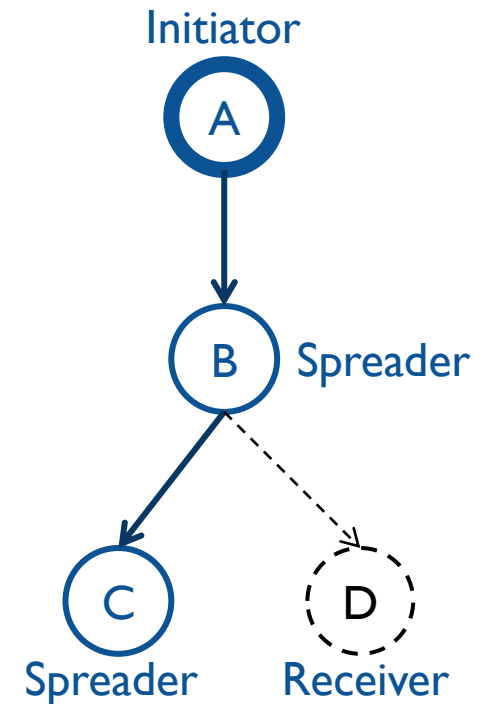
T	User	Tweet content
1	A	Check this: http://www.example.com/
2	B	http://www.example.com/ is interesting



Modeling Information Cascades

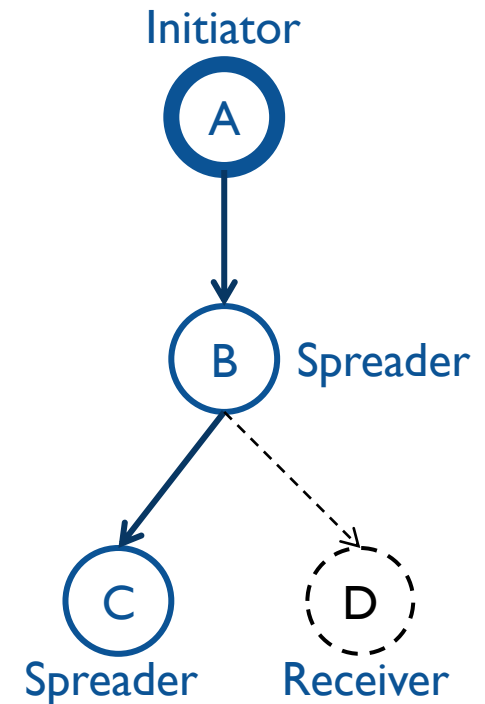
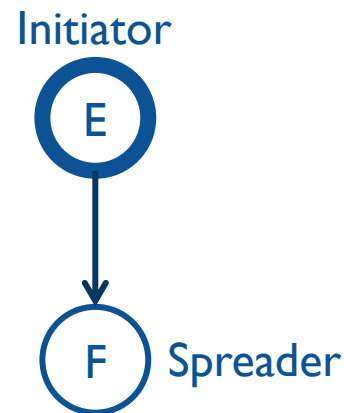
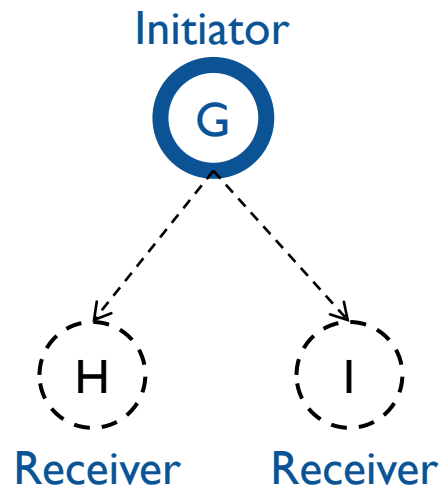
► Hierarchical tree model

T	User	Tweet content
1	A	Check this: http://www.example.com/
2	B	http://www.example.com/ is interesting
3	C	Interesting link: http://www.example.com/



Modeling Information Cascades

- ▶ Hierarchical tree model
 - ▶ URL propagation pattern is a forest

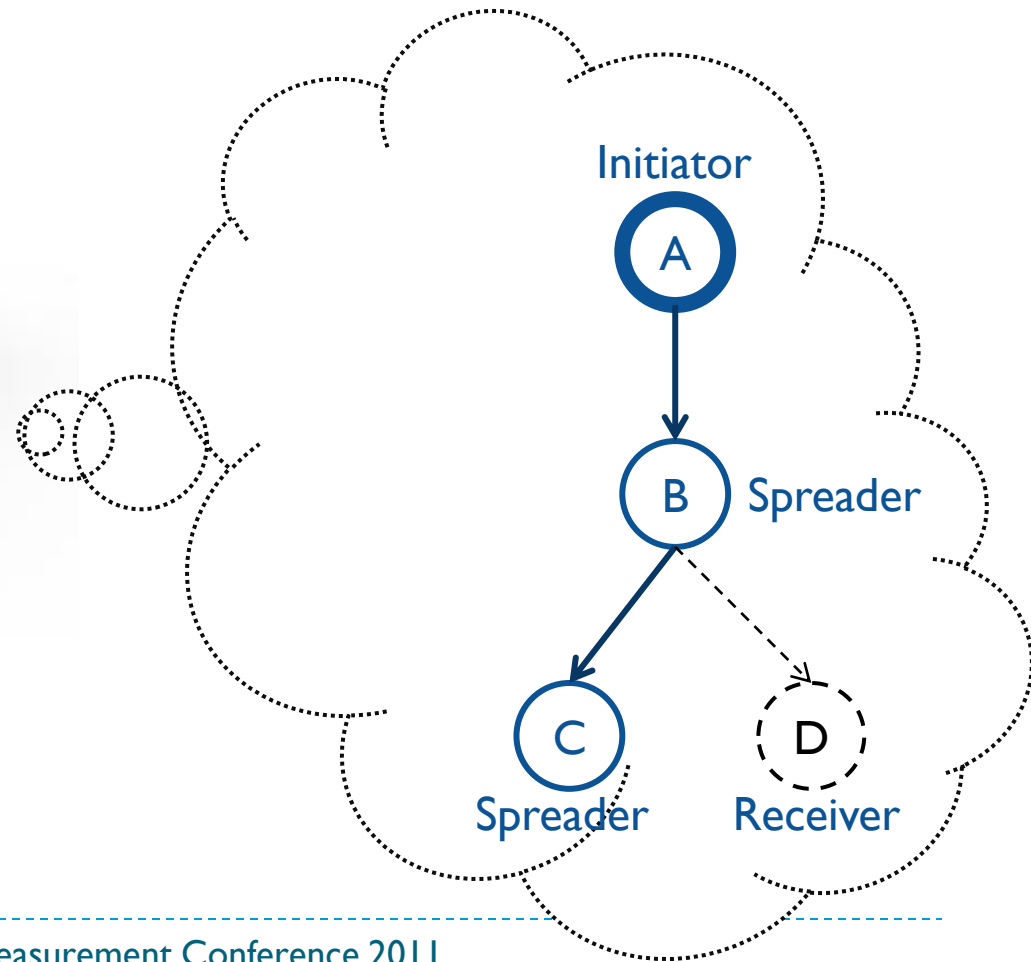


Modeling Information Cascades

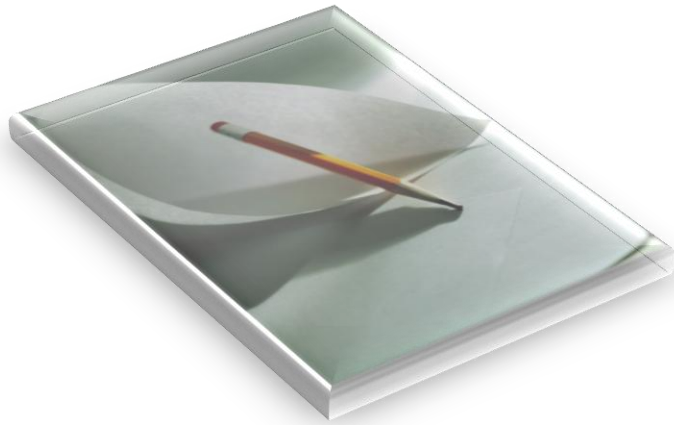
► Hierarchical tree model



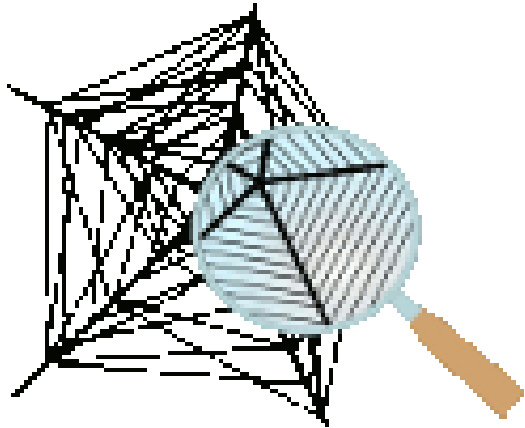
Audience



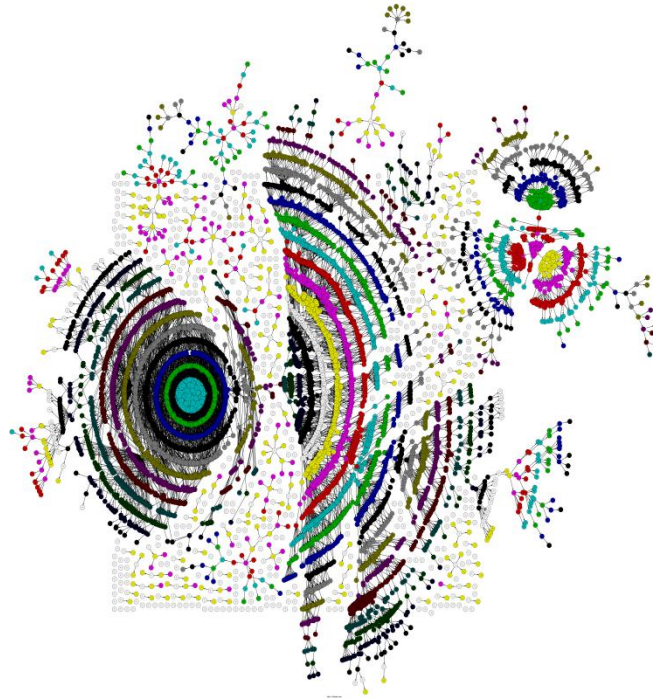
Results



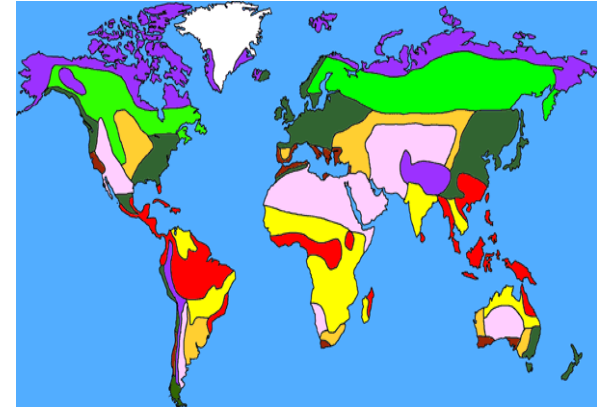
Groups of Analysis



Content



Shape



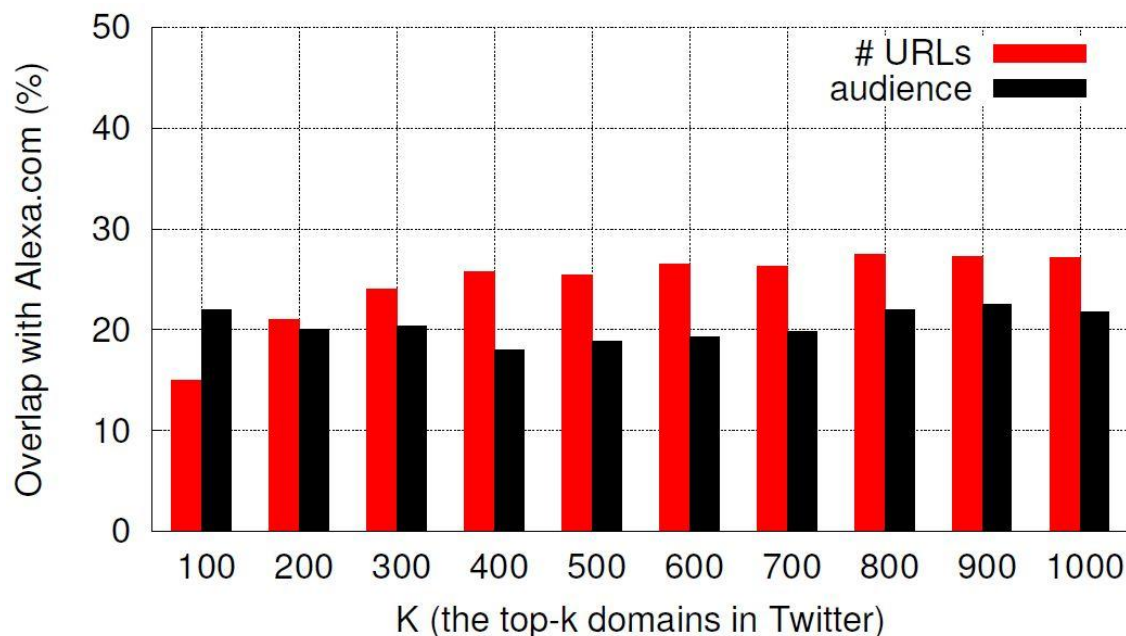
Geography

Content Analysis

- ▶ What URLs are popularly shared on Twitter? Do they come from the popular domains in the Web?
- ▶ Does all content, including those published by unpopular domains, benefit from Word-of-Mouth?
- ▶ What kinds of content is popular in Word-of-Mouth discovery of the Web?

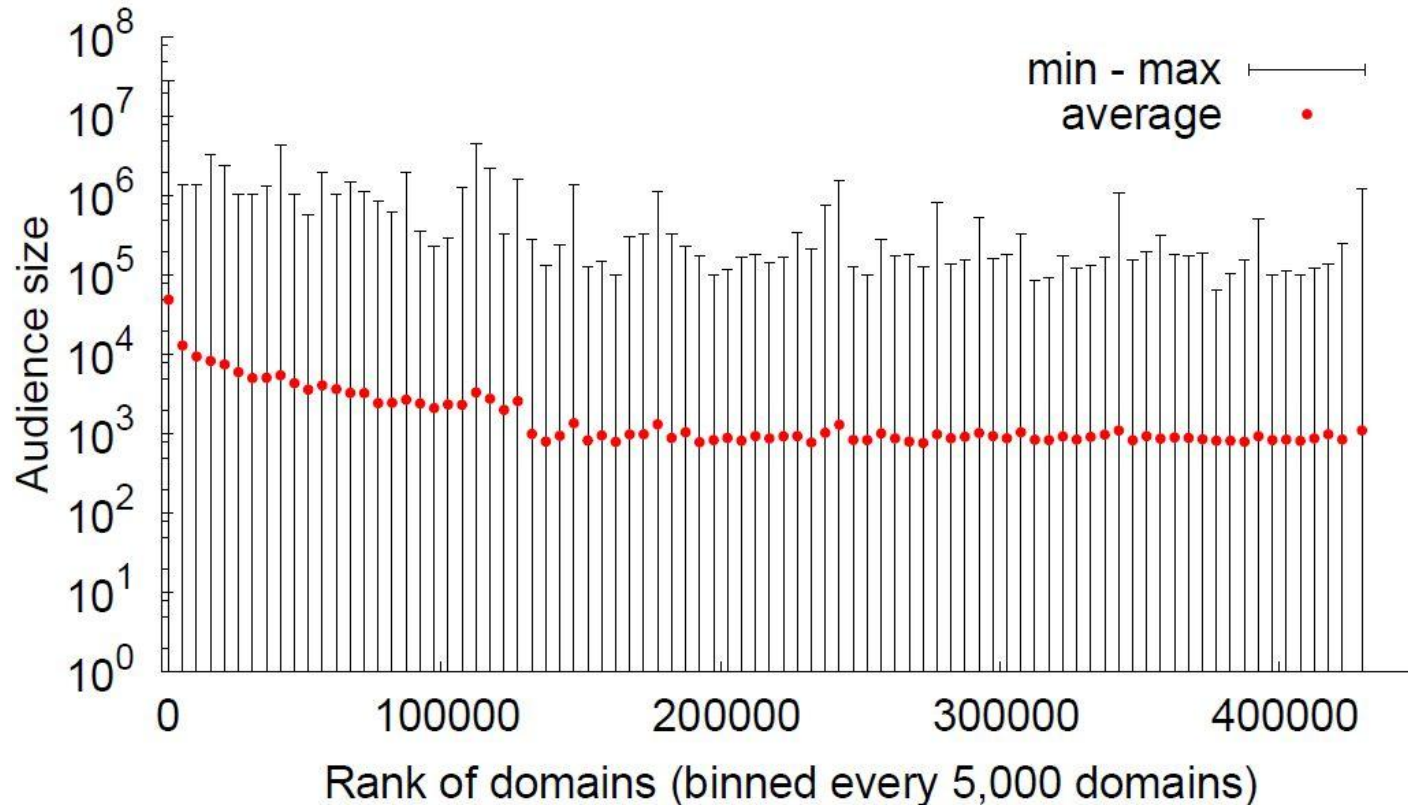
What URLs are popularly shared on Twitter? Do they come from the popular domains in the Web?

Rank	Top list	Description	URLs	Alexa rank
1	twitpic.com	photo sharing	8.5%	103
2	blip.fm	music sharing	3.0%	6,736
3	youtube.com	video sharing	2.1%	3
4	plurk.com	social journal	2.1%	1,146
5	tumblr.com	blog	1.4%	100



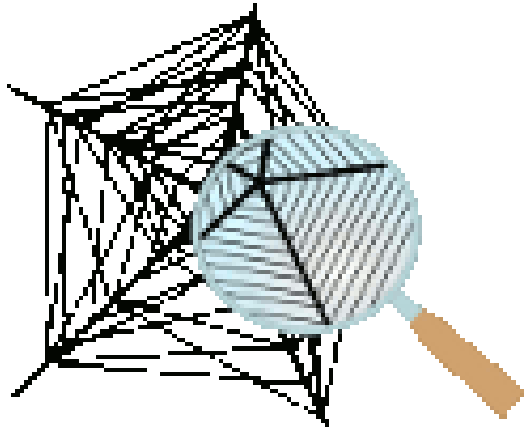
- ▶ Word-of-mouth can help popularize niche content

Does all content, including those published by unpopular domains, benefit from Word-of-Mouth?

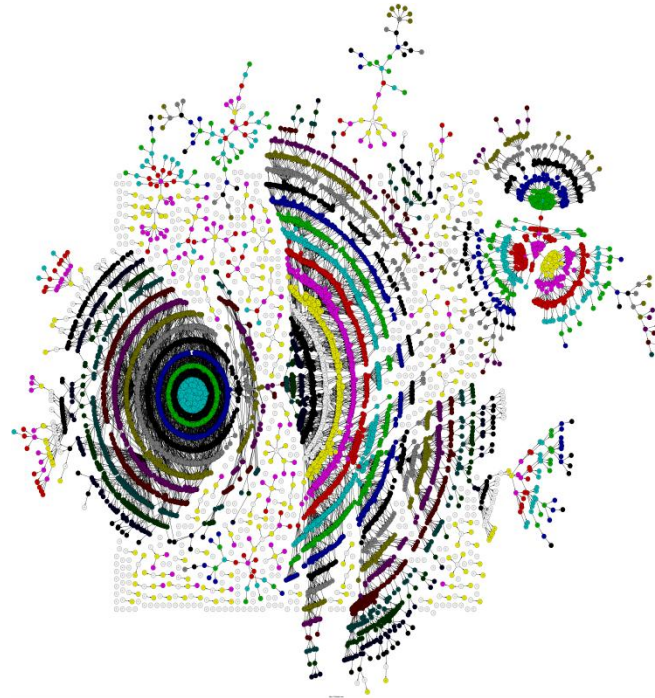


- ▶ Word-of-mouth gives all URLs and content (both popular and non-popular) a chance to become popular

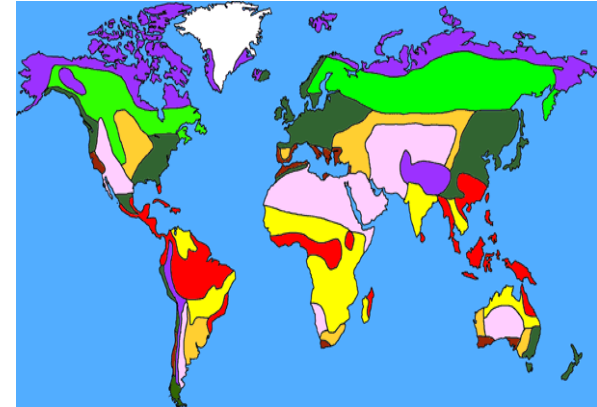
Groups of Analysis



Content



Shape



Geography

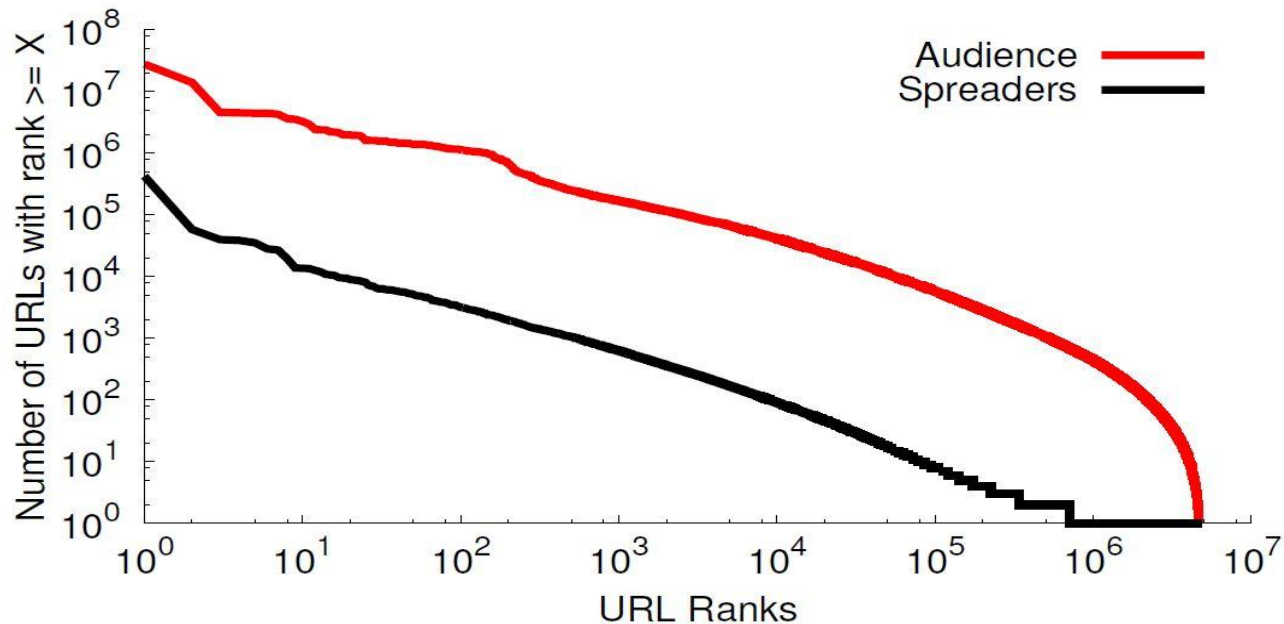
The Shape of Word-of-Mouth

- ▶ How large is the largest Word-of-Mouth?
- ▶ What are the typical structures of Word-of-Mouth propagation trees?
- ▶ Is having multiple initiators essential for yielding a large cascade?

How large is the largest Word-of-Mouth?

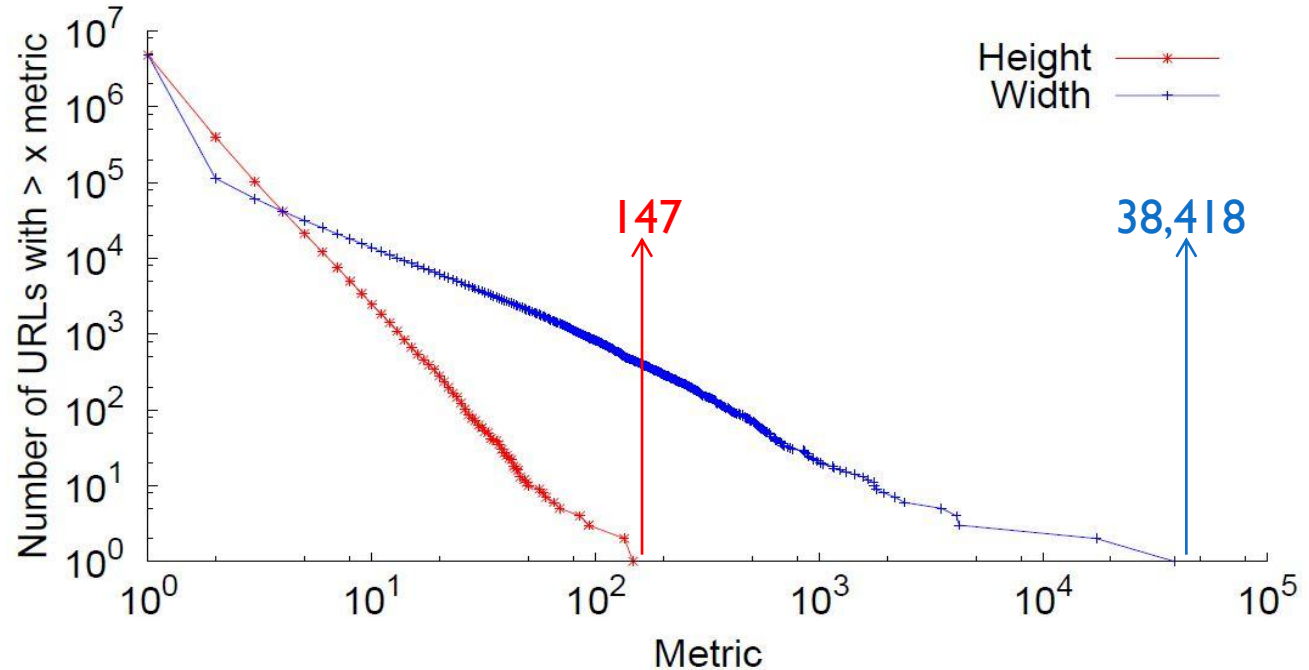
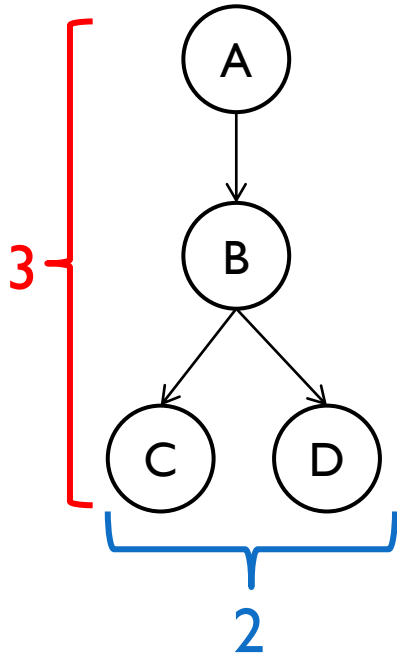
▶ URL popularity

- ▶ Most popular: 426,820 spreaders and audience of 28M users
- ▶ Average: 3 spreaders and audience of 843 users



- ▶ Word-of-mouth can incur extremely large cascades

What are the typical structures of Word-of-Mouth propagation trees?



- ▶ Cascade trees are much wider than they are deep
 - ▶ 0.1% of the trees have width > 20
 - ▶ 0.005% of the trees have height > 20

Twitter Cascades vs E-mail Cascades

- ▶ D. Liben-Nowell and J. Kleinberg

- ▶ Tracing Information Flow on a Global Scale using Internet Chain-Letter Data, PNAS, 2008

Data Source	Nodes	Height	Width
PNAS [31]	18,119	288 (med)	82
Twitter	26,227	23 (max)	17,255
PNAS (estimated)	980	16 (med)	4
Twitter (900-1,100)	980 ± 0.02	20 ± 0.25	398 ± 0.27
PNAS (estimated)	162	3 (med)	1
Twitter (100-300)	162 ± 0.02	10 ± 0.03	86 ± 0.03

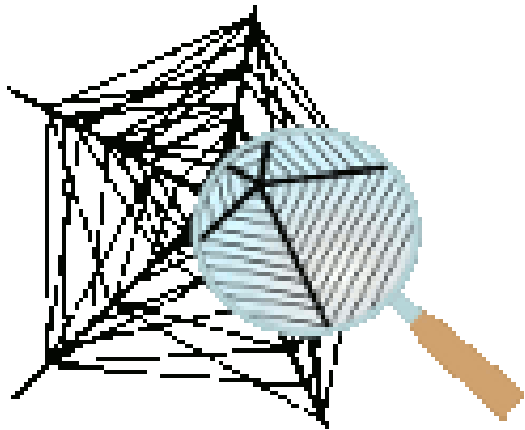


Twitter

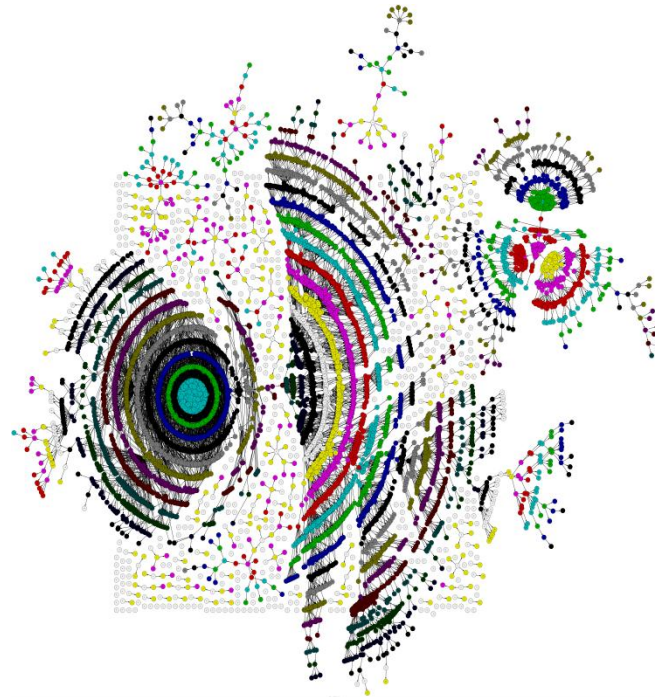


e-mail

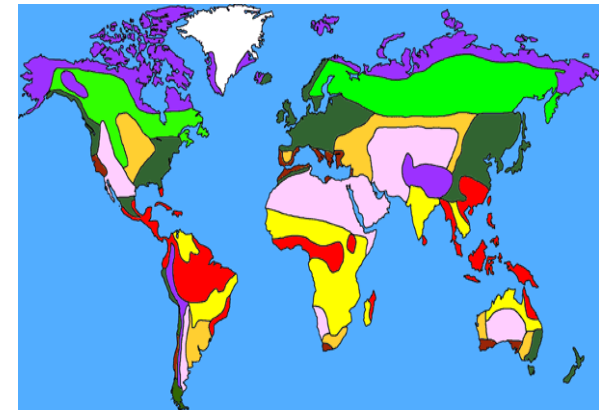
Groups of Analysis



Content



Shape

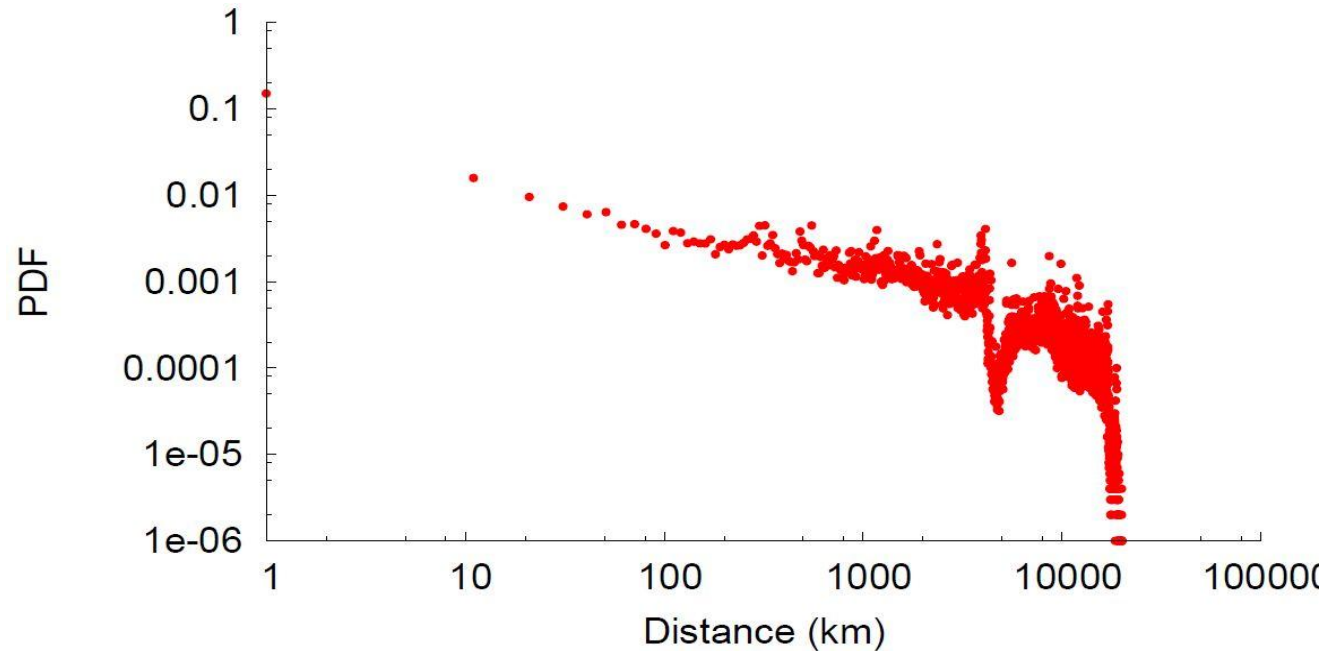
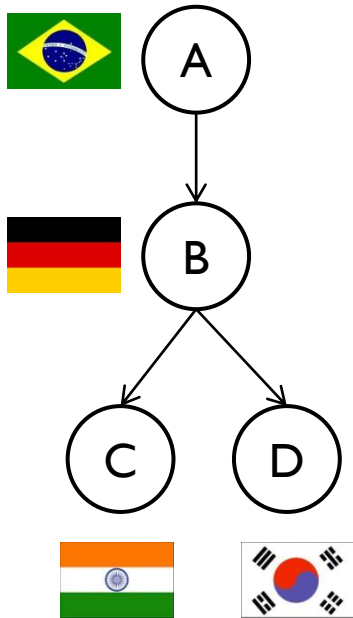


Geography

Content Distribution

- ▶ How geographically distributed are the propagation trees?
- ▶ Do friend-of-friends help in making the content reach distant locations?
- ▶ Is the level of locality dependent on the content type or the audience of the content producer?

How geographically distributed are the propagation trees?



- ▶ Users within a short geographical distance have a higher probability of posting the same URL

Conclusions

- ▶ First large scale analysis of Word-of-Mouth Web based content discovery
 - ▶ All contents have a chance to reach a large audience
 - ▶ Propagation trees on Twitter are wide and shallow
 - ▶ Advertising
 - ▶ Content is consumed locally
 - ▶ Caching design and recommendation
 - ▶ Strong correlation between the audience size and the number of initiators
 - ▶ Content produced by users with a small audience is usually consumed by users located within a small physical distance

Questions



▶ **tiagorm@dcc.ufmg.br**