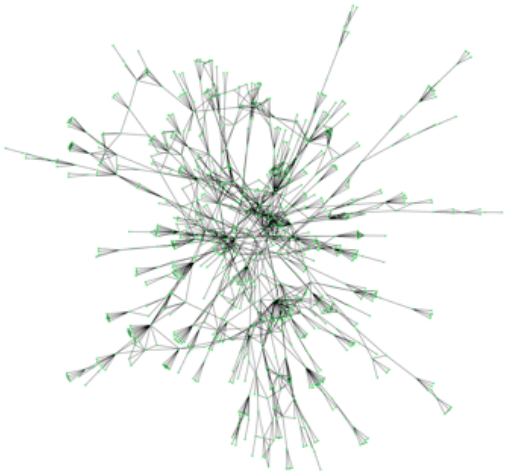
Carnegie Mellon

Part 3: Case Studies



Outline

- Part 1: How do networks form, evolve, collapse?
- Part 2: What tools can we use to study networks?
- Part 3: Case studies
  - How do ideas diffuse through a network?
  - How to detect communities?
  - How do we detect anomalies in networks?

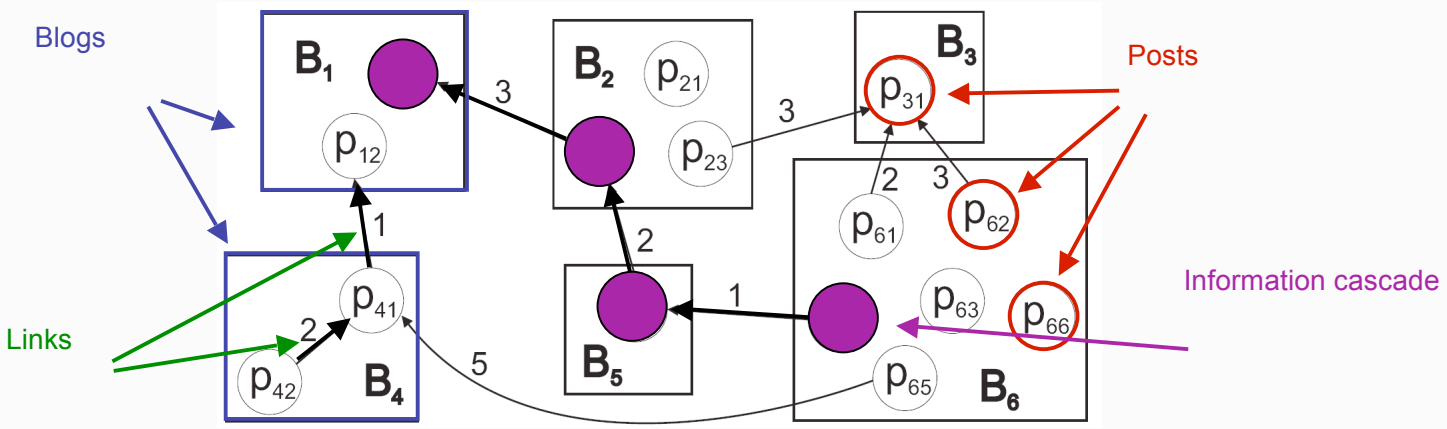
Carnegie Mellon

Part 3: Case Studies

- Q4: How do ideas diffuse through a network?
  - Cascades
  - Epidemiological modeling of cascades
  - Outbreak detection
- Q5: How can we extract communities?
  - Using PCA on structure
  - Factorization
- Q6: What sort of anomaly detection can we perform?
  - Fraud detection on E-bay
  - Spam detection

Carnegie Mellon

Cascading Behavior in Large Blog Graphs



How does information propagate over the blogosphere?

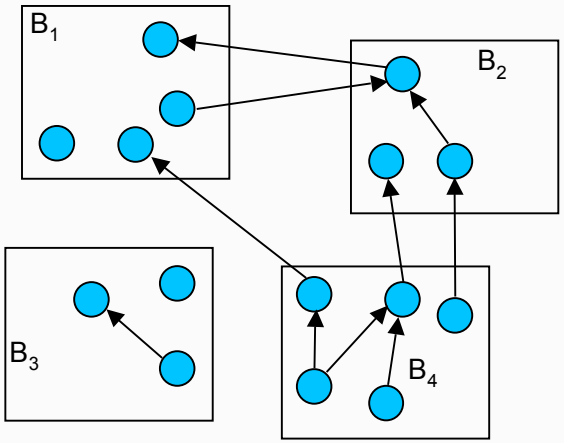
J. Leskovec, M.McGlohon, C. Faloutsos, N. Glance, M. Hurst. Cascading Behavior in Large Blog Graphs. SDM 2007.

4

Immediate Goals

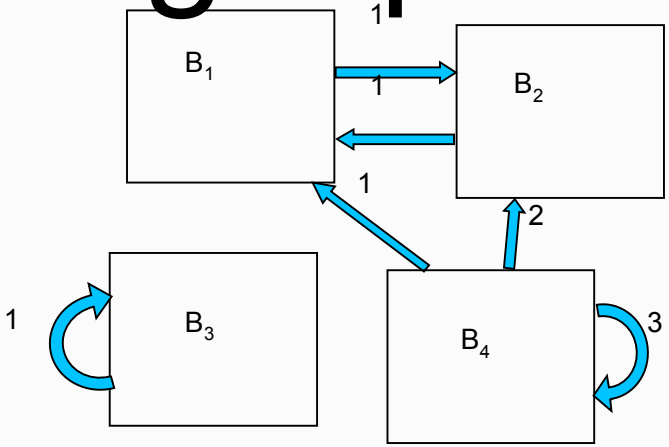
- Temporal questions: Does popularity have half-life?
- Topological questions: What topological patterns do posts and blogs follow? What shapes to cascades take on? Stars? Chains? Something else?
- Models: Can we build a generative model that mimics properties of cascades?

Cascades on the Blogosphere



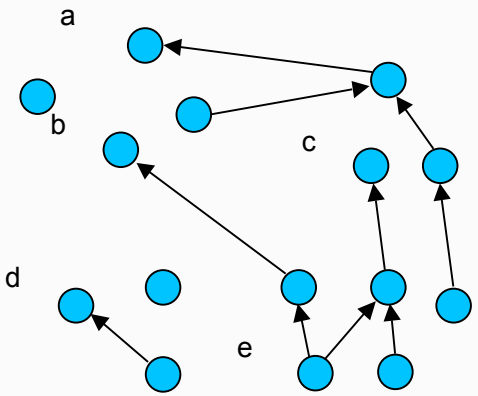
Blogosphere

blogs + posts



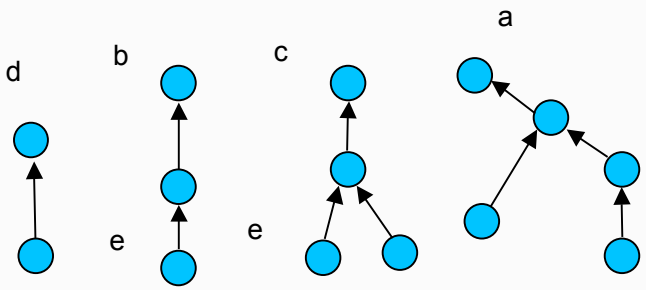
Blog network

links among blogs



Post network

links among posts



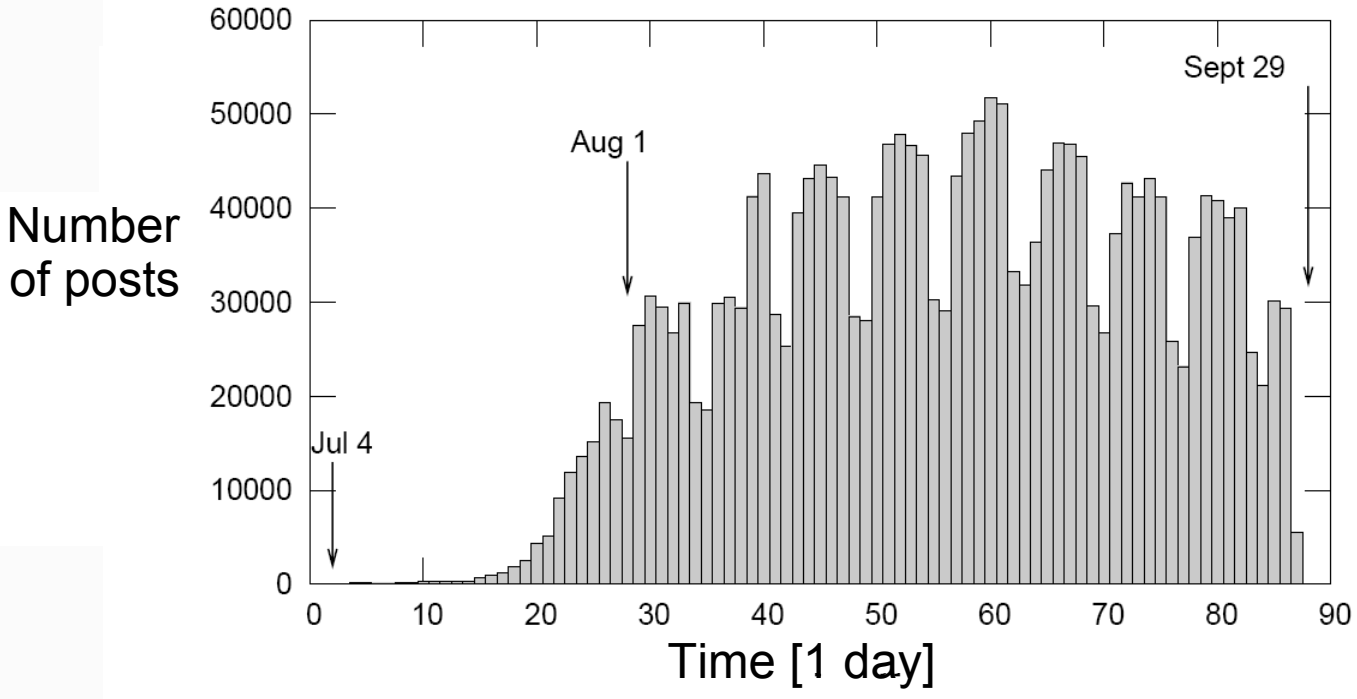
Cascades

Cascade is graph induced by a time ordered propagation of information (edges)

Carnegie Mellon

Blog data

- 45,000 blogs participating in cascades
- All their posts for 3 months (Aug-Sept '05)
- 2.4 million posts
- ~5 million links (245,404 inside the dataset)

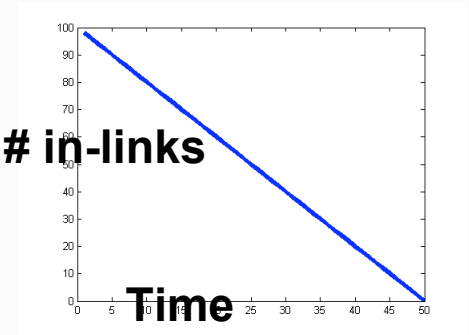


Carnegie Mellon

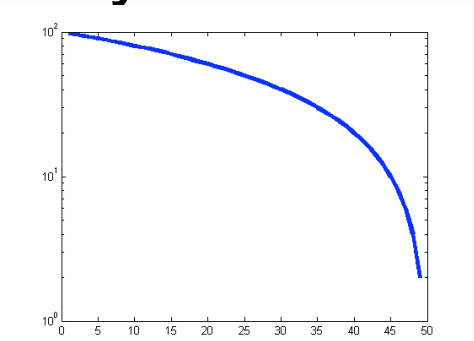
Temporal Observations

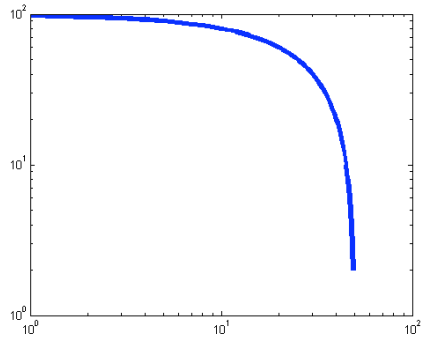
How does post popularity change over time?

- Does popularity decay at a constant rate?

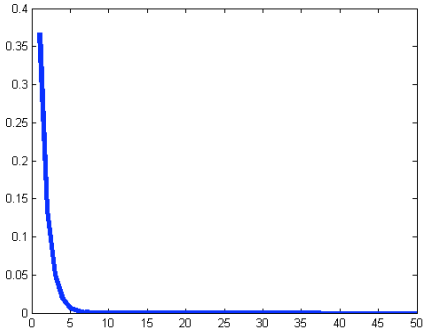


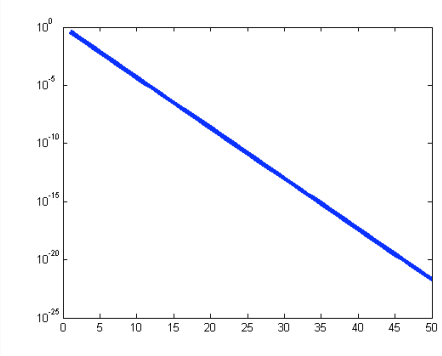
Linear-linear scale

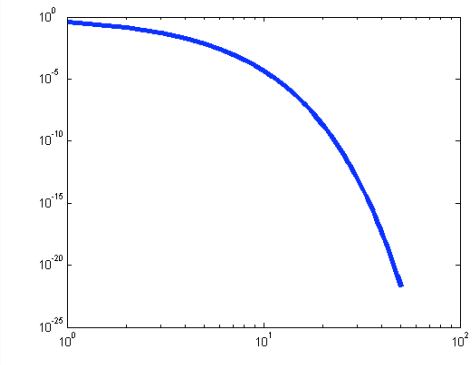

Log-linear scale


Log-log scale

- With an exponential ("half life")?





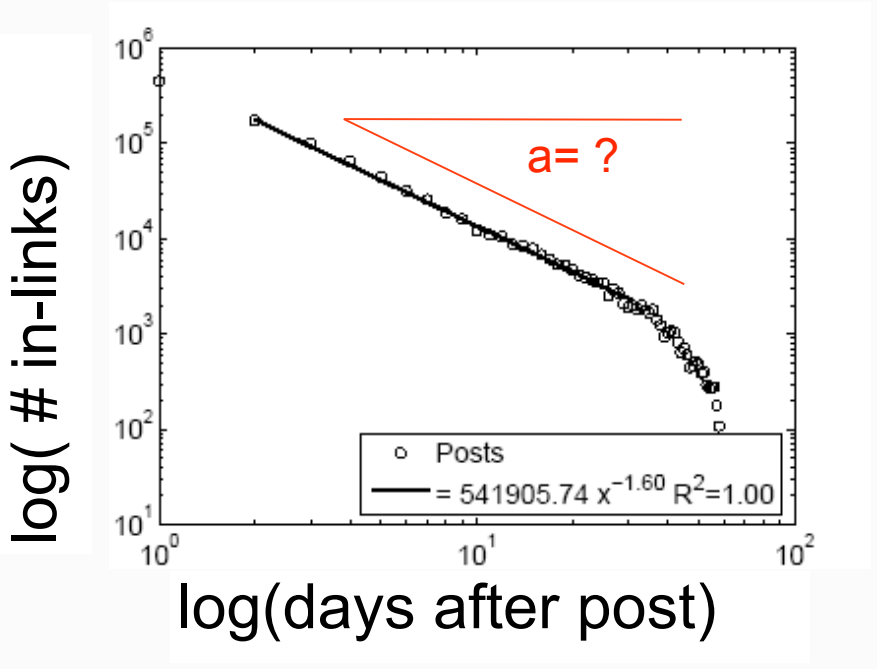


Carnegie Mellon

Temporal Observations

How does post popularity change over time?

Post popularity dropoff follows a power law...



Carnegie Mellon

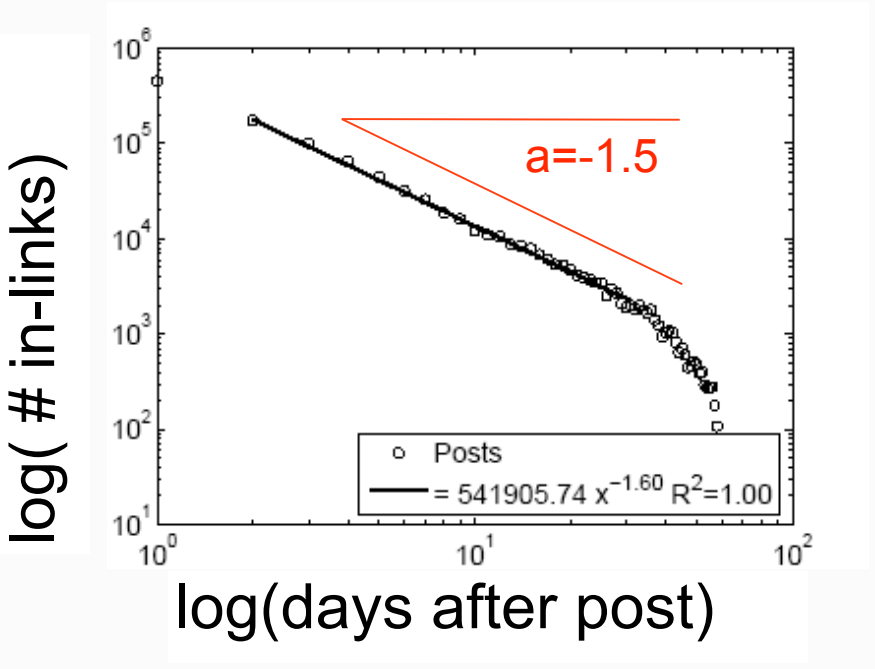
Temporal Observations

How does post popularity change over time?

Post popularity dropoff follows a power law identical to that found in communication response times in [Vazquez+2006].

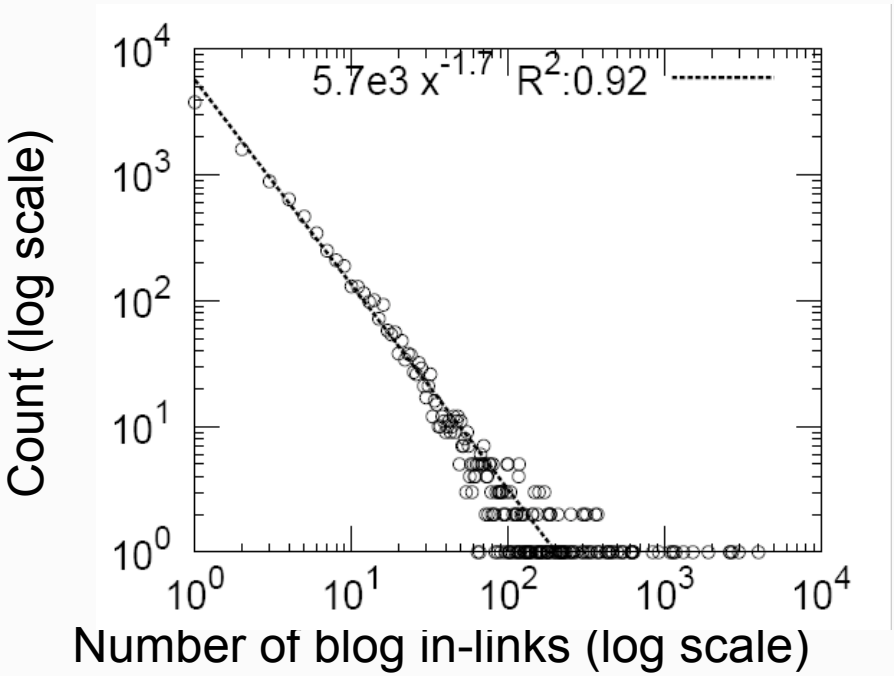
Observation 1: *The probability that a post written at time t_p acquires a link at time $t_p + \Delta$ is:*

$p(t_p+\Delta) \propto \Delta^{-1.5}$



What is topology of blogs?

44,356 nodes, 122,153 edges. Half of blogs belong to largest connected component.



In- and out-degree follow power law distribution. In-degree exponent -1.7, out-degree exponent -3.

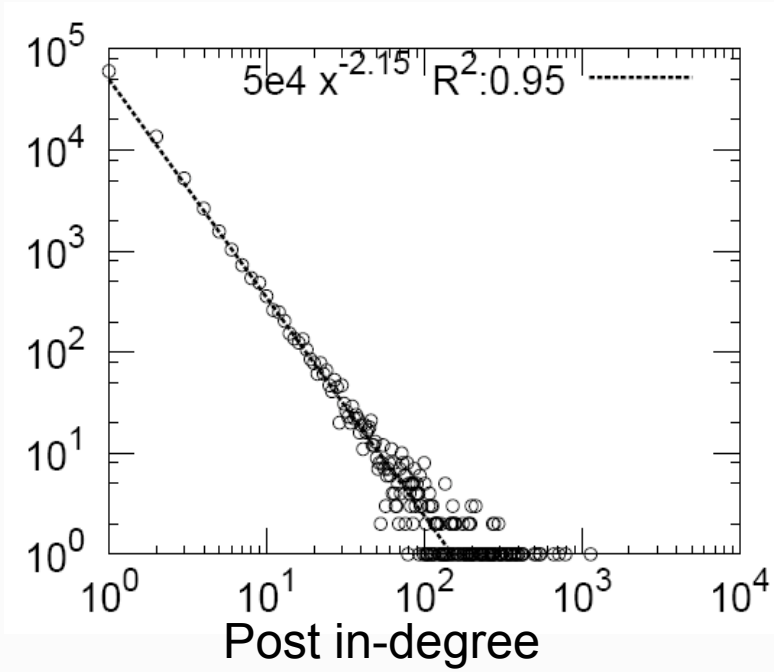
Strong rich-get-richer phenomena.

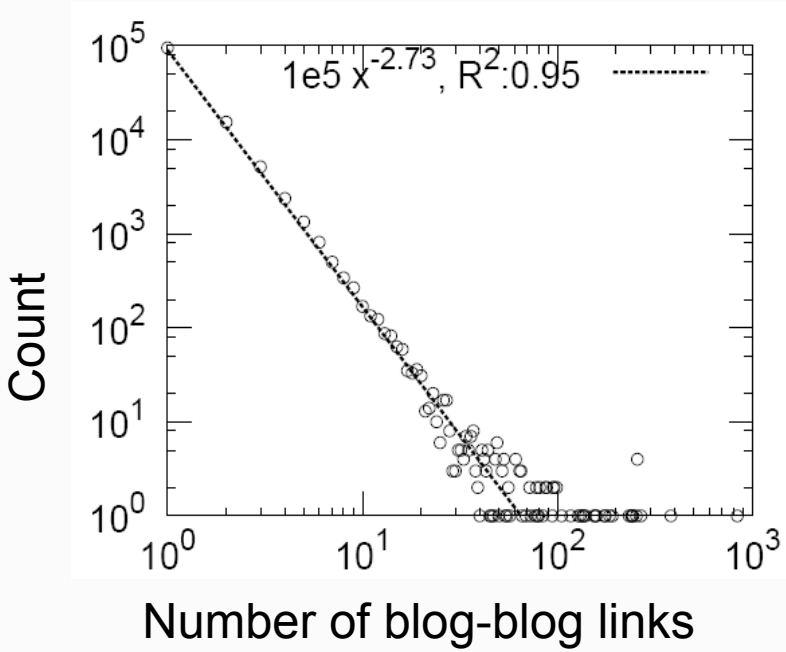
Carnegie Mellon

Post Network

2.4M nodes, 250K edges

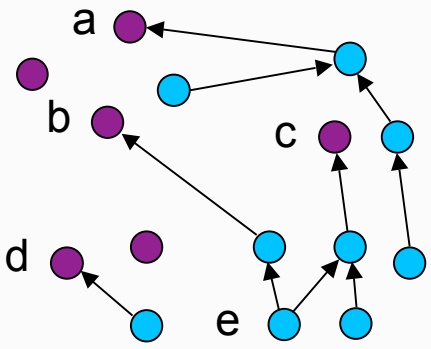
Both in- and out-degree follow power laws. In-degree exponent -2.1, out-degree exponent -3.

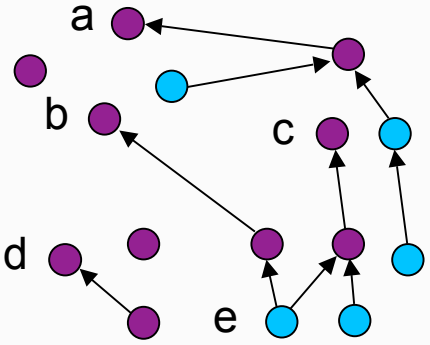


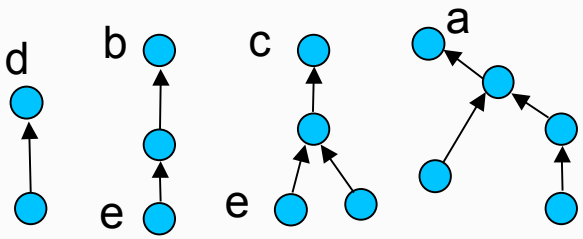


Topological patterns: Cascades

- Procedure for gathering cascades:
  - Find all initiators (nodes with out-degree 0)
  - Follow in-links
  - Produces directed acyclic graph
  - Count cascade shapes (use our multi-level graph isomorphism testing algorithm)



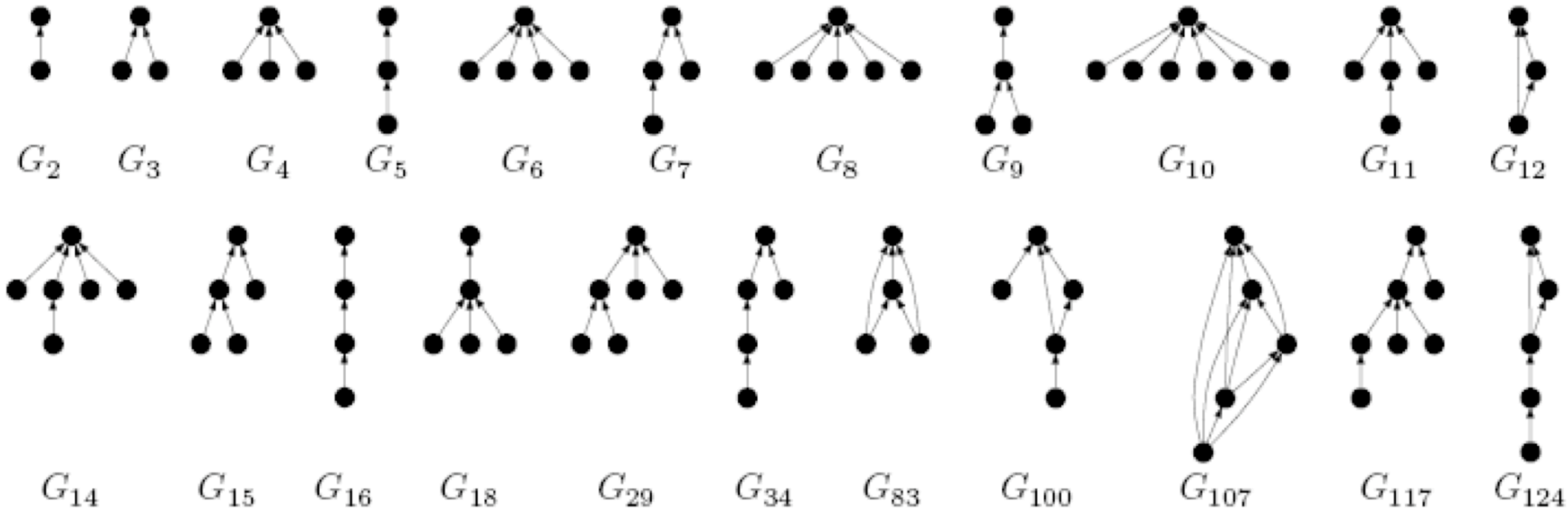




Topological Observations

How do we measure how information flows through the network?

Common cascade shapes extracted using algorithms in [Leskovec, Singh, Kleinberg; PAKDD 2006].



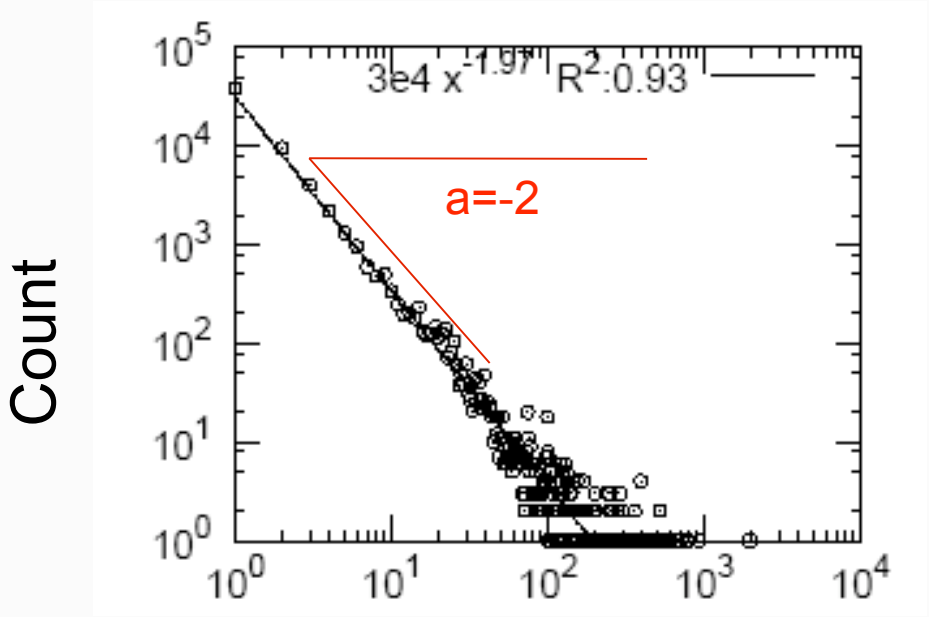
Topological Observations

What graph properties do cascades exhibit?

Cascade size distributions also follow power law.

Observation 2: *The probability of observing a cascade on n nodes follows a Zipf distribution:*

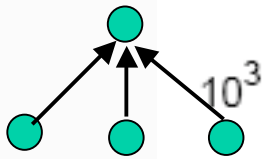
$p(n) \propto n^{-2}$

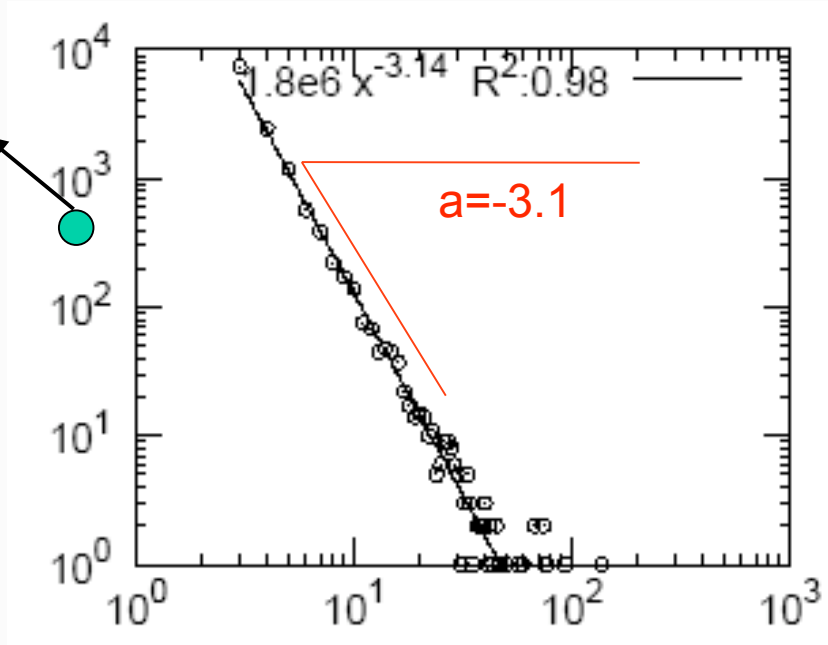


Topological Observations

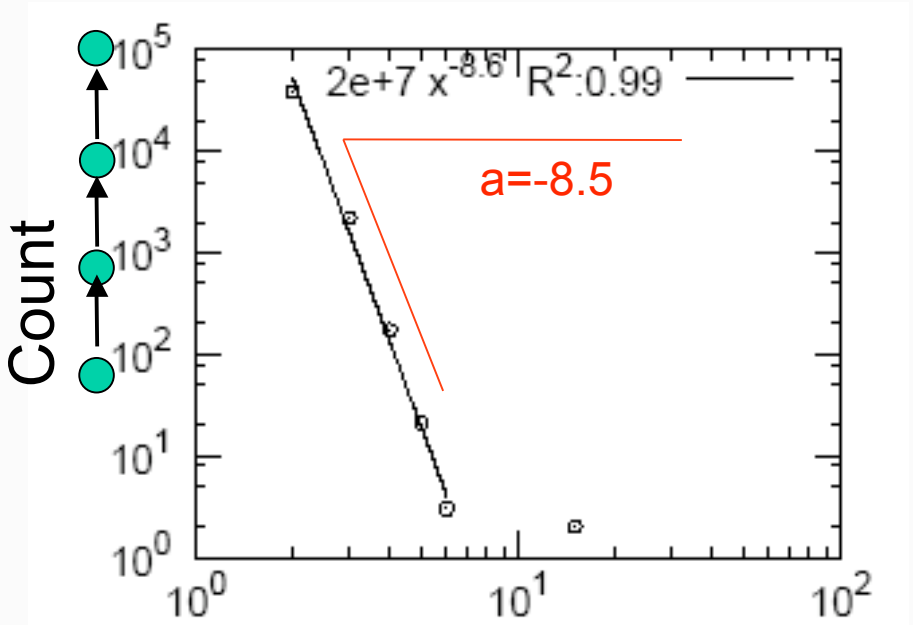
What graph properties do cascades exhibit?

Stars and chains also follow a power law, with different exponents (star -3.1, chain -8.5).





Size of star (# nodes)



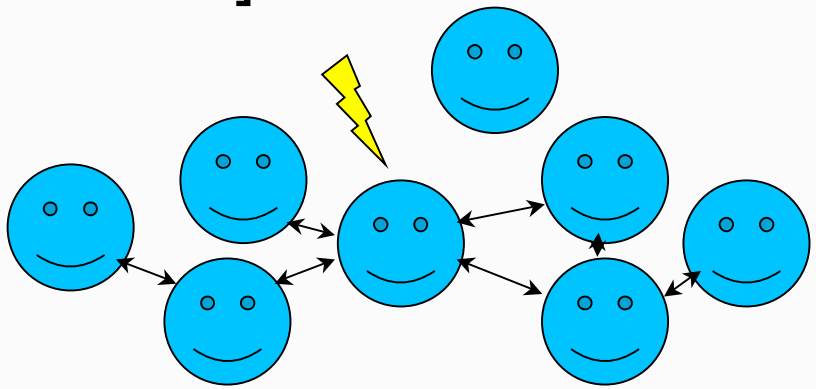
Size of chain (# nodes)

Epidemiological models

- We consider modeling cascade generation as an epidemic, with ideas as viruses.
- We use the SIS (flu-like) model:
  - At any time, an entity is in one of two states: susceptible or infected.
  - One parameter β determines how easily spreading conversations are.
  - [Hethcote2000]

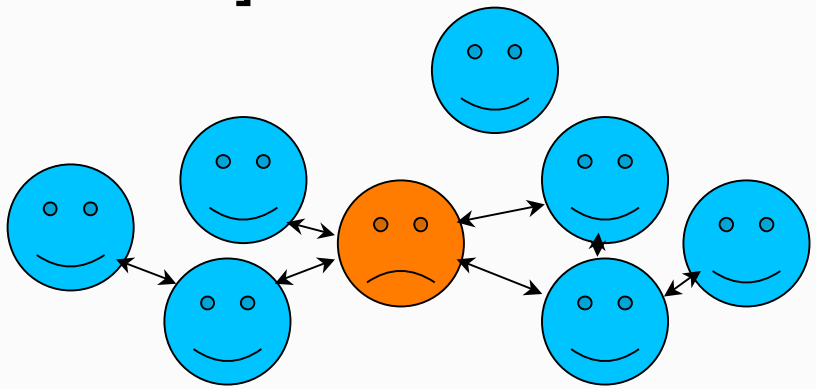
Epidemiological models

- We consider modeling cascade generation as an epidemic, with ideas as viruses.
- We use the SIS (flu-like) model:
  - At any time, an entity is in one of two states: susceptible or infected.
  - One parameter β determines how easily spreading conversations are.
  - [Hethcote2000]



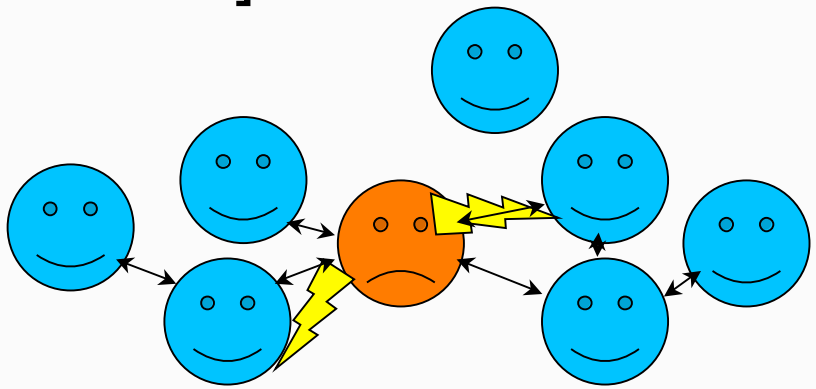
Epidemiological models

- We consider modeling cascade generation as an epidemic, with ideas as viruses.
- We use the SIS (flu-like) model:
  - At any time, an entity is in one of two states: susceptible or infected.
  - One parameter β determines how easily spreading conversations are.
  - [Hethcote2000]



Epidemiological models

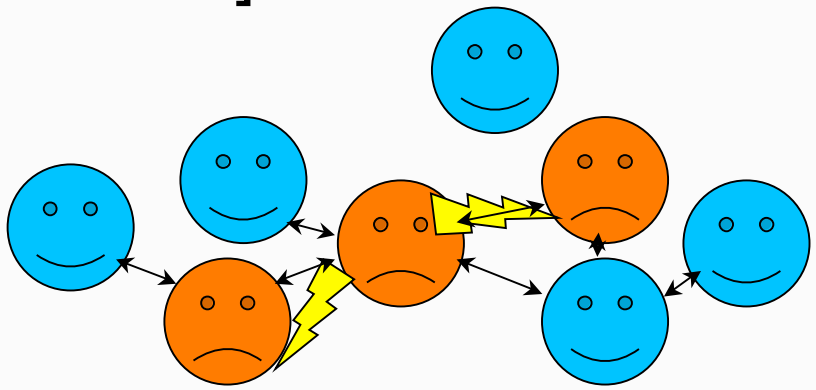
- We consider modeling cascade generation as an epidemic, with ideas as viruses.
- We use the SIS (flu-like) model:
  - At any time, an entity is in one of two states: susceptible or infected.
  - One parameter β determines how easily spreading conversations are.
  - [Hethcote2000]



Carnegie Mellon

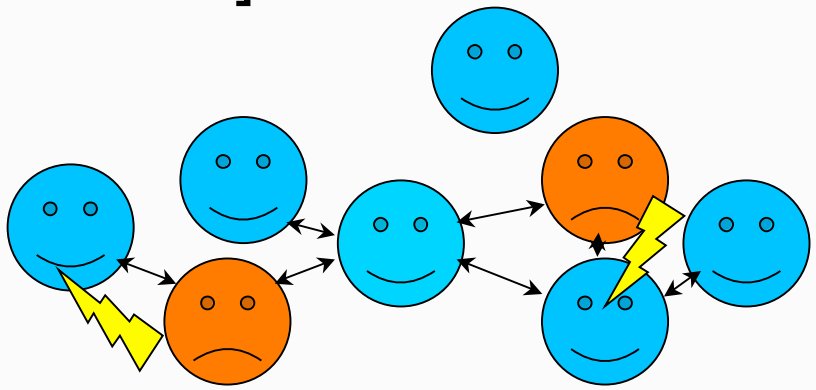
Epidemiological models

- We consider modeling cascade generation as an epidemic, with ideas as viruses.
- We use the SIS (flu-like) model:
  - At any time, an entity is in one of two states: susceptible or infected.
  - One parameter β determines how easily spreading conversations are.
  - [Hethcote2000]



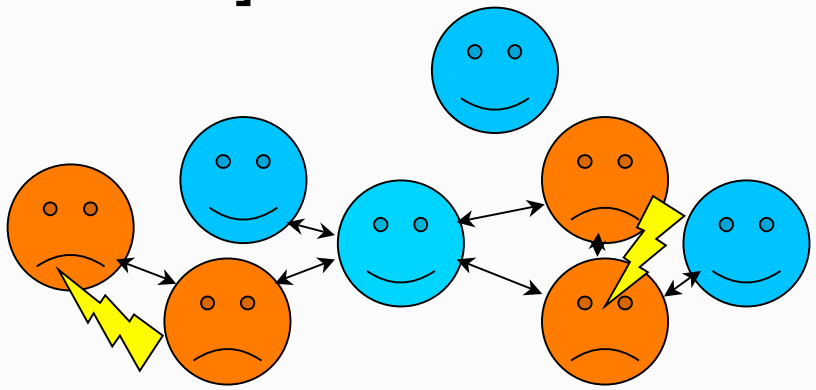
Epidemiological models

- We consider modeling cascade generation as an epidemic, with ideas as viruses.
- We use the SIS (flu-like) model:
  - At any time, an entity is in one of two states: susceptible or infected.
  - One parameter β determines how easily spreading conversations are.
  - [Hethcote2000]



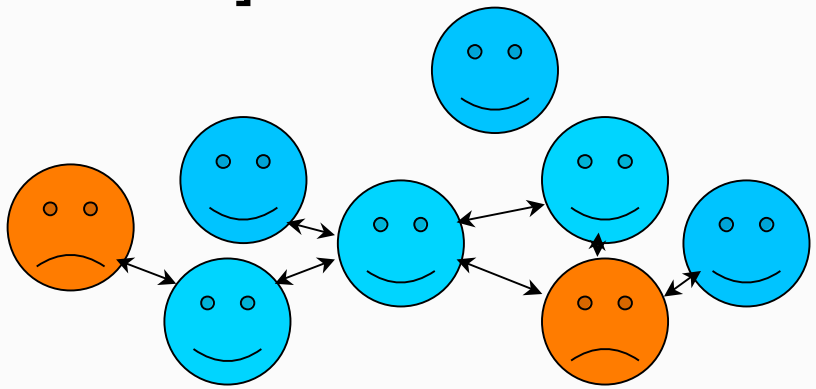
Epidemiological models

- We consider modeling cascade generation as an epidemic, with ideas as viruses.
- We use the SIS (flu-like) model:
  - At any time, an entity is in one of two states: susceptible or infected.
  - One parameter β determines how easily spreading conversations are.
  - [Hethcote2000]



Epidemiological models

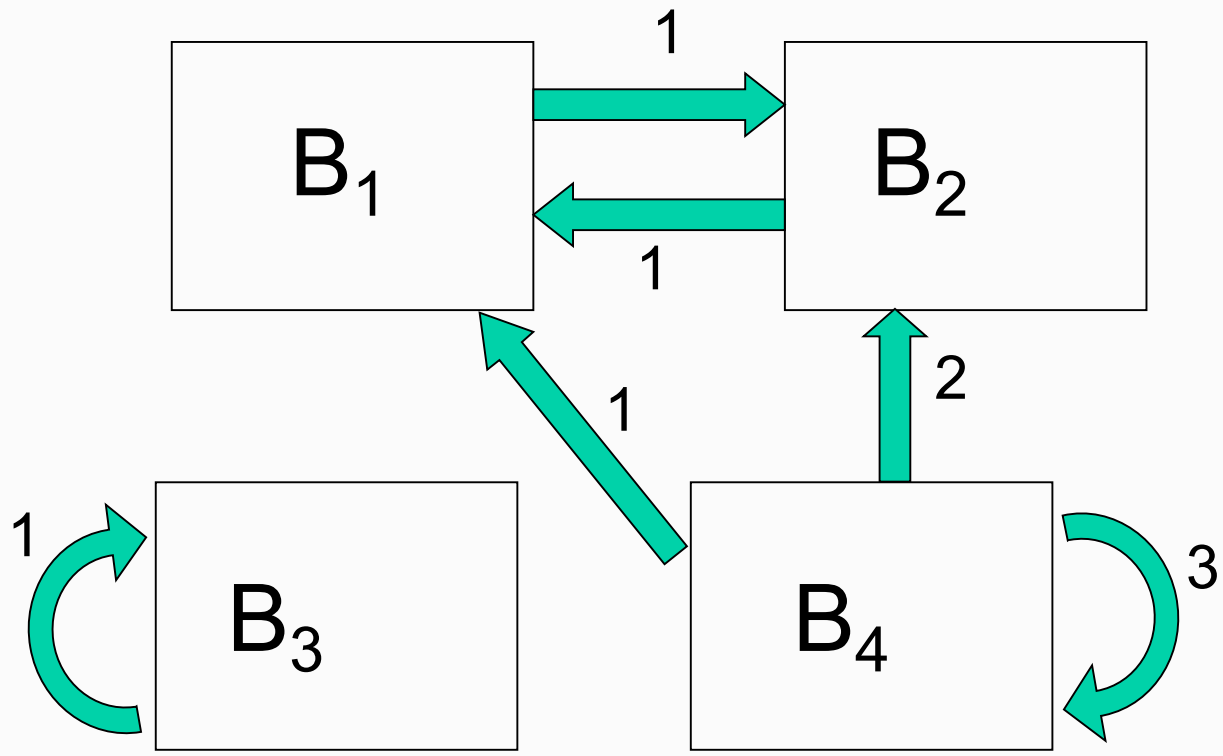
- We consider modeling cascade generation as an epidemic, with ideas as viruses.
- We use the SIS (flu-like) model:
  - At any time, an entity is in one of two states: susceptible or infected.
  - One parameter β determines how easily spreading conversations are.
  - [Hethcote2000]



Carnegie Mellon

Cascade Generation Model

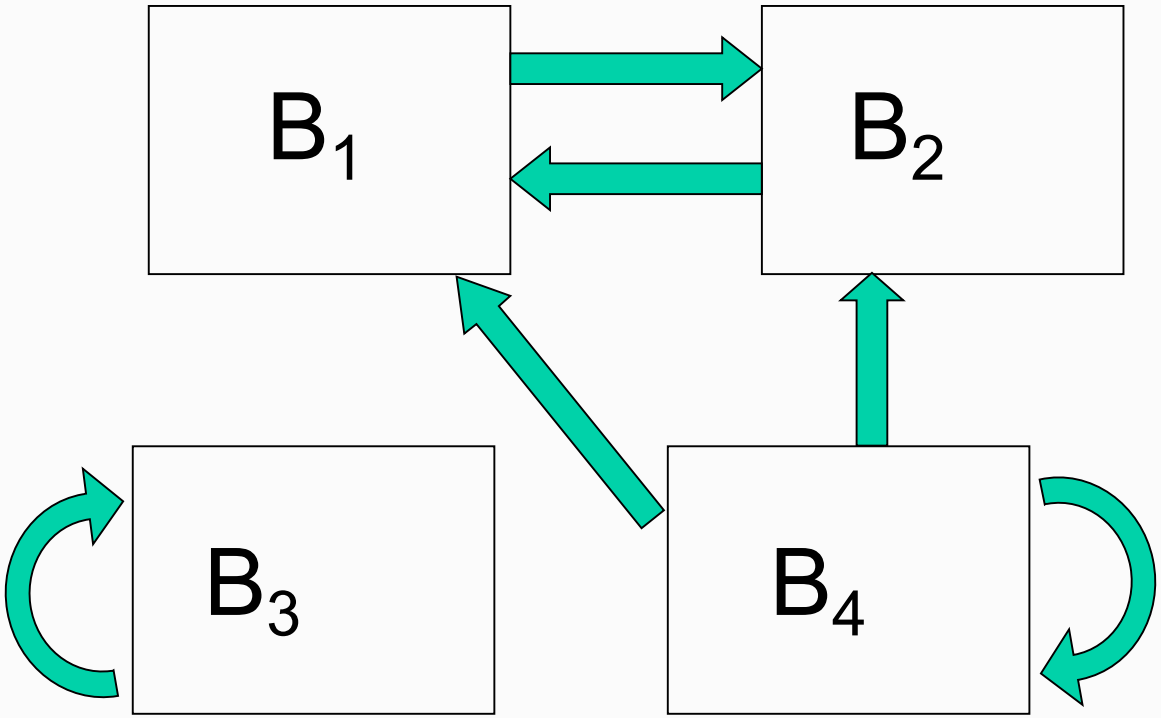
0. Begin with Blog Net.



Carnegie Mellon

Cascade Generation Model

0. Begin with Blog Net, but ignore edge weights.



Example–

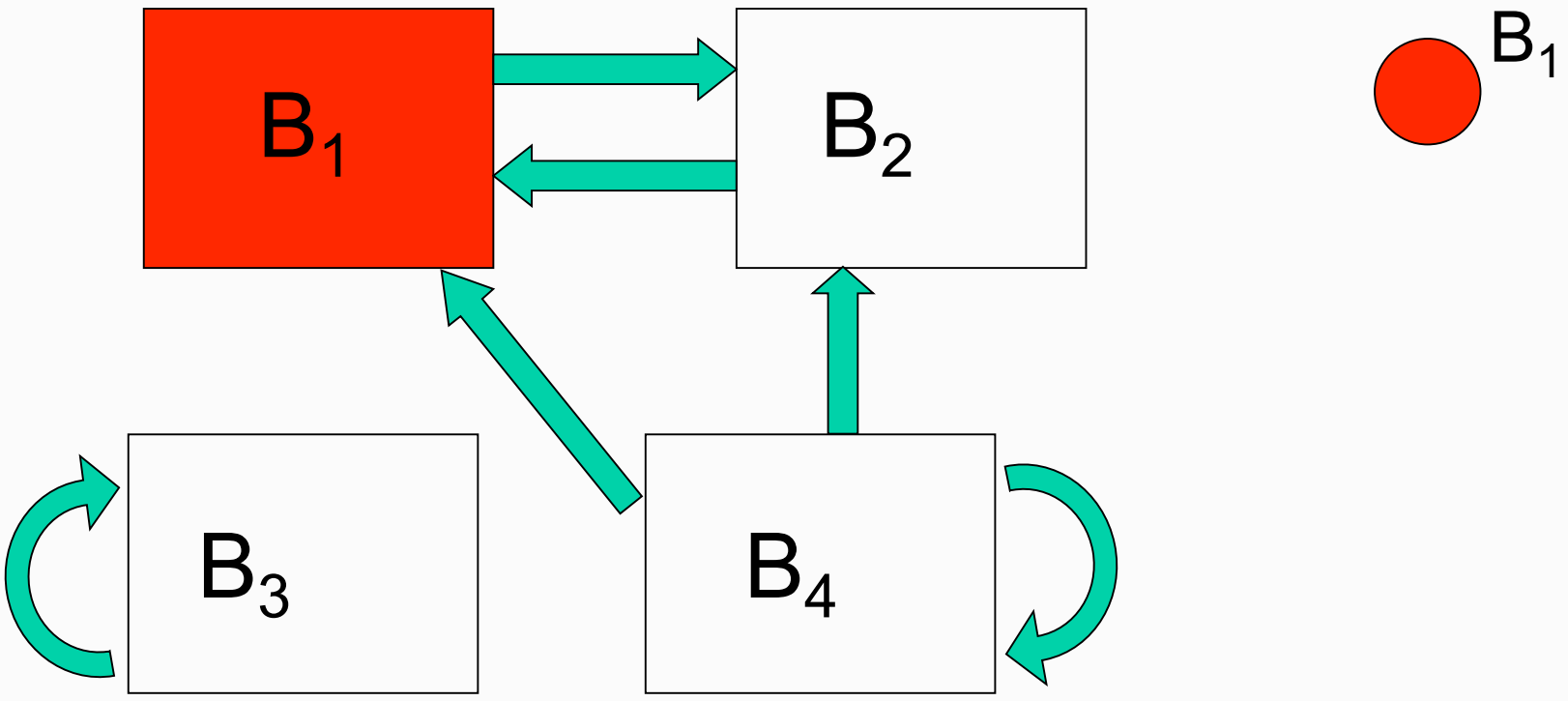
B1 links to B2, B2 links to B1, B4 links to B2 and B1, as well as itself

B3 is isolated, linking to itself.

Carnegie Mellon

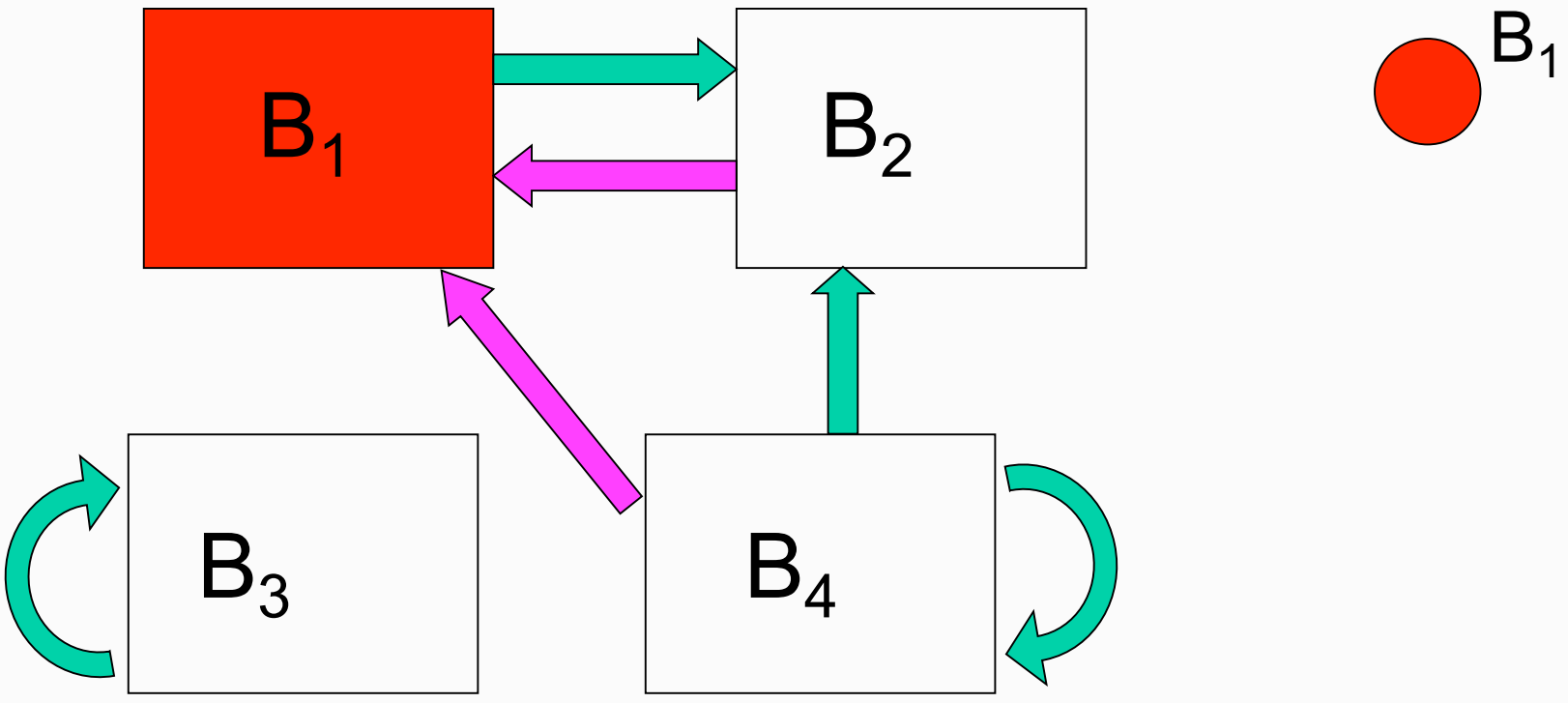
Cascade Generation Model

1. Randomly pick a blog to infect, add node to cascade



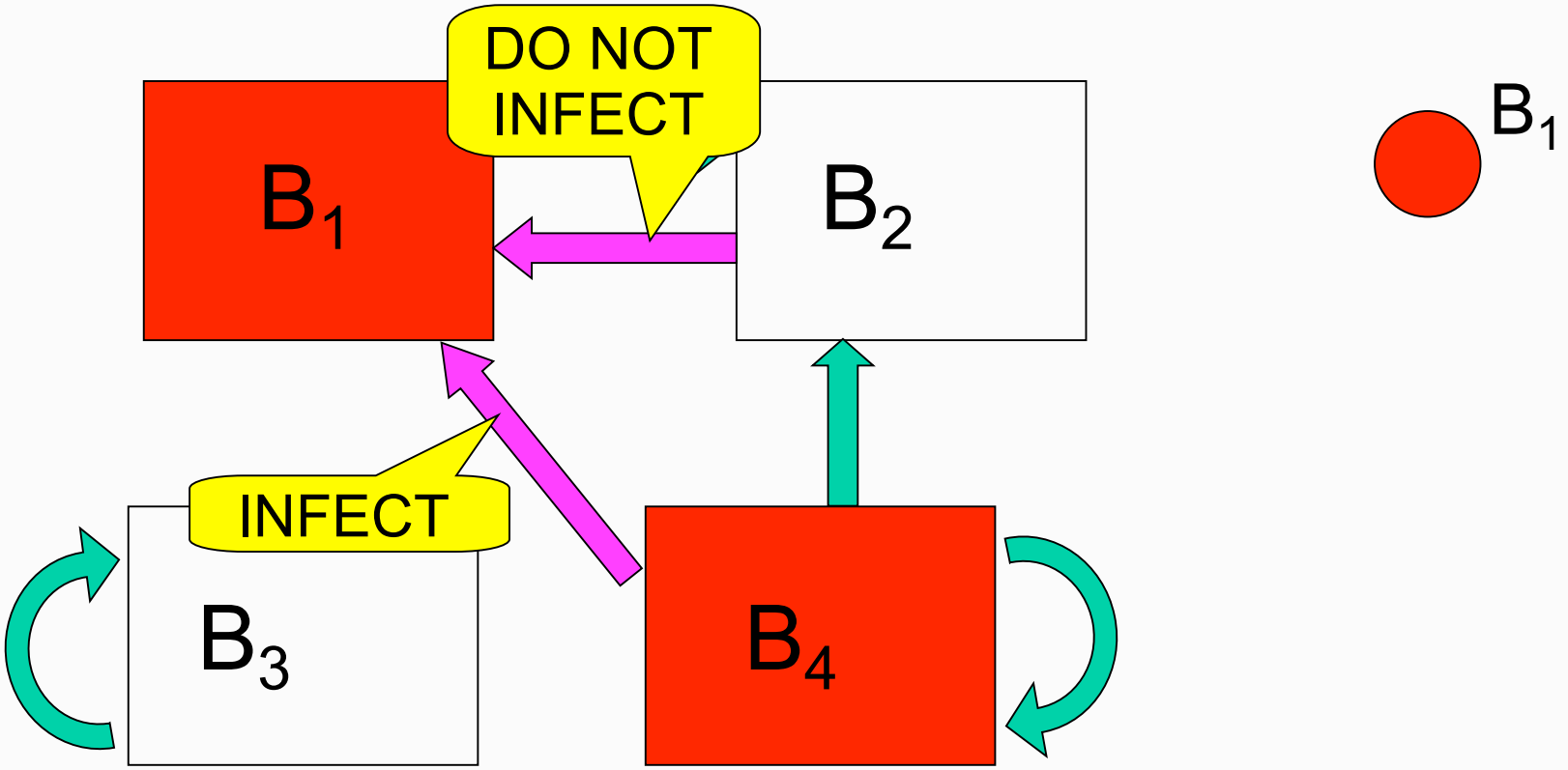
Cascade Generation Model

2. Infect each in-linked neighbor with probability β.



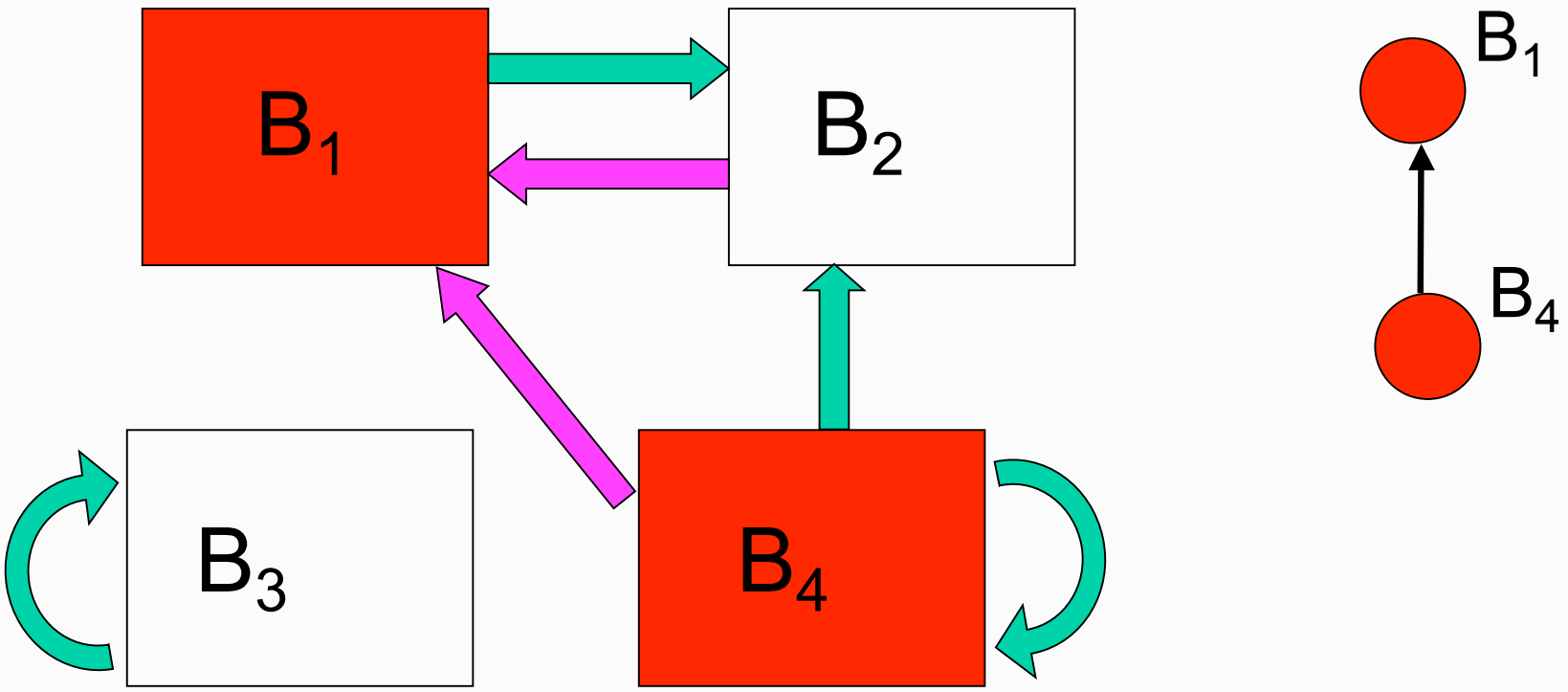
Cascade Generation Model

2. Infect each in-linked neighbor with probability β.



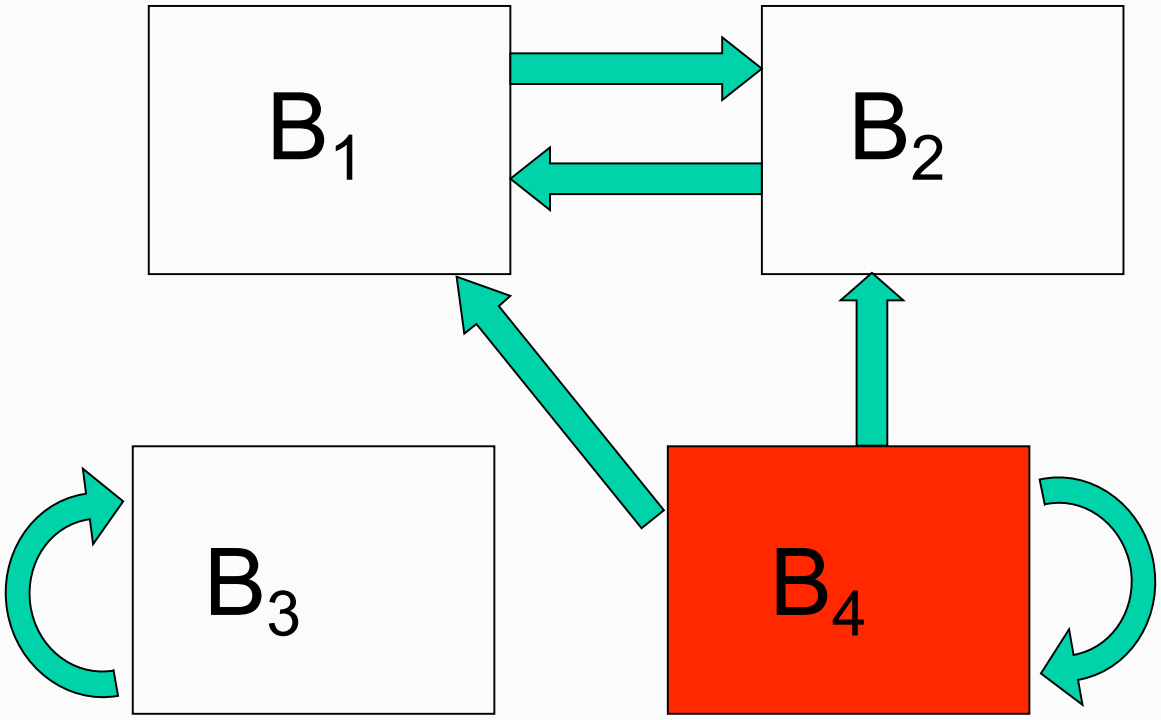
Cascade Generation Model

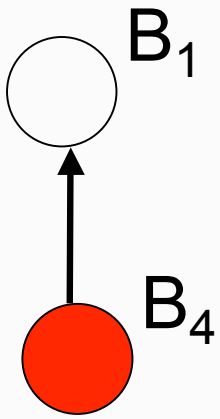
3. Add infected neighbors to cascade.



Cascade Generation Model

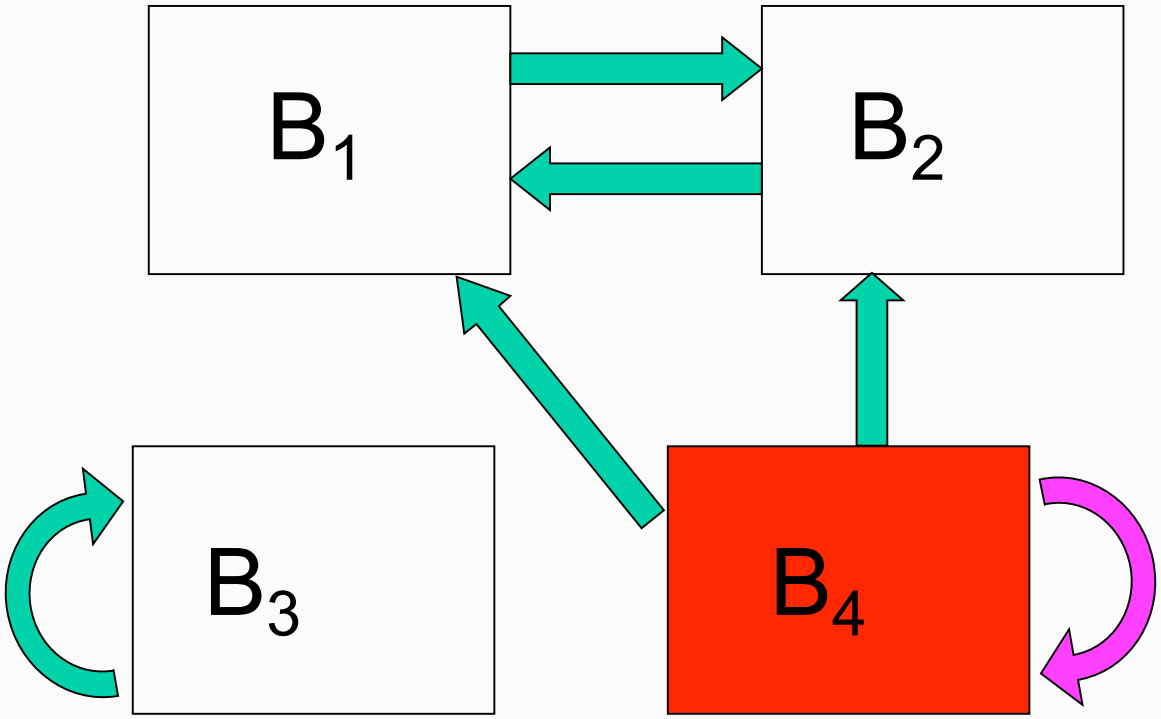
4. Set “old” infected nodes to uninfected.

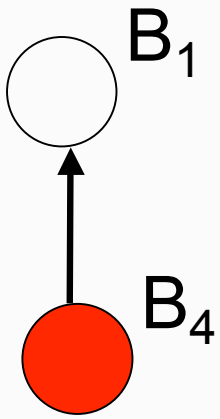




Cascade Generation Model

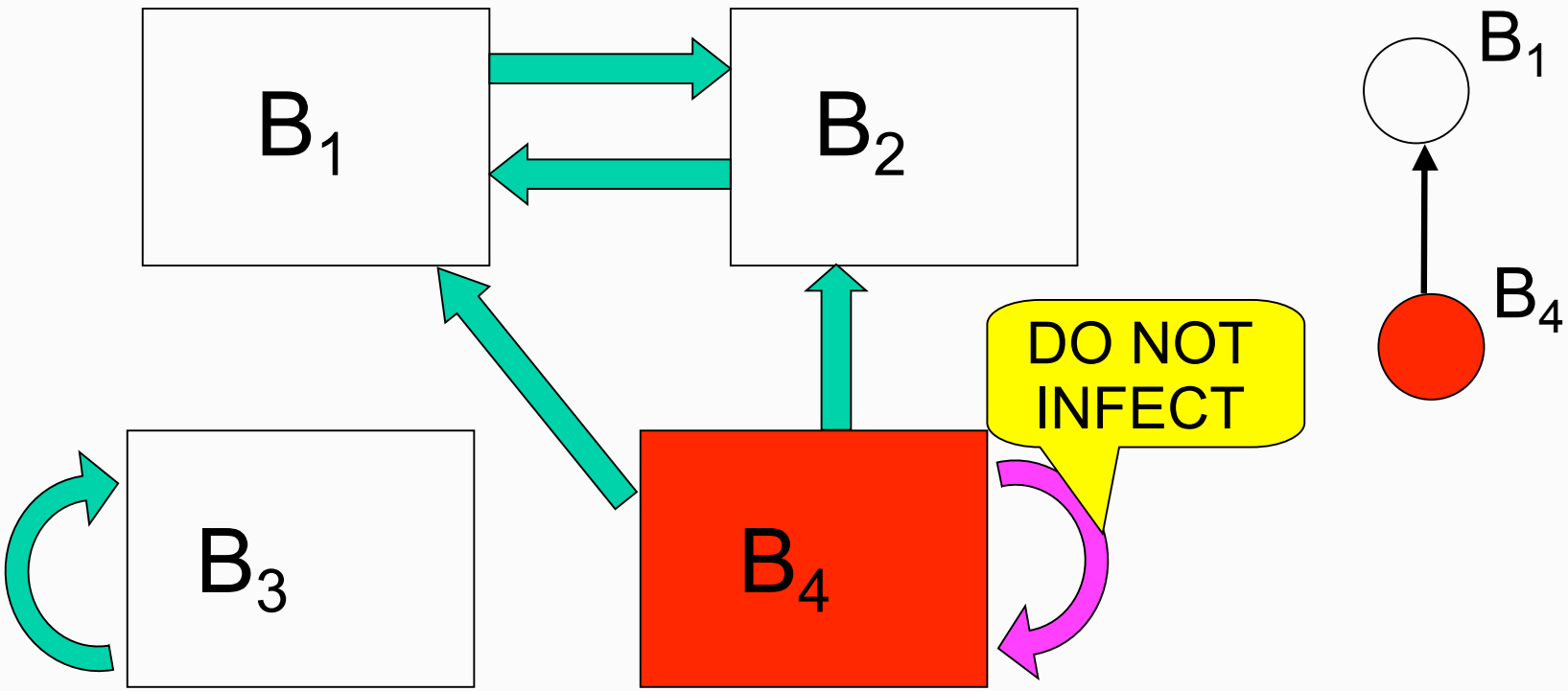
4. Set “old” infected nodes to uninfected. Repeat steps 2-4 until no nodes are infected.





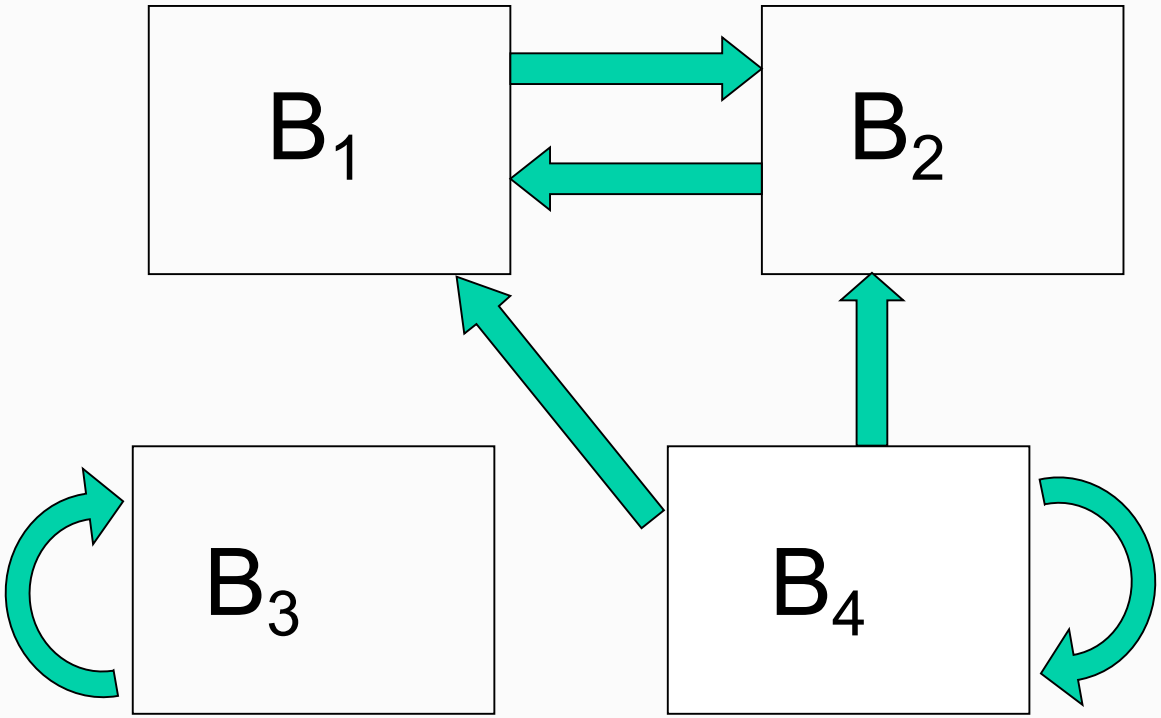
Cascade Generation Model

4. Set “old” infected nodes to uninfected. Repeat steps 2-4 until no nodes are infected.

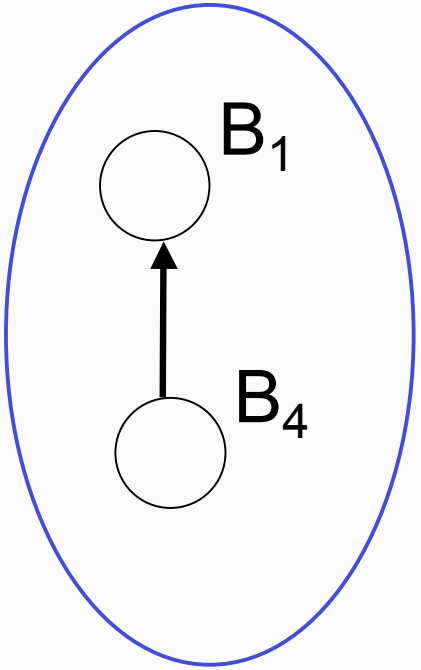


Cascade Generation Model

4. Set "old" infected nodes to uninfected. Repeat steps 2-4 until no nodes are infected.



Completed cascade!

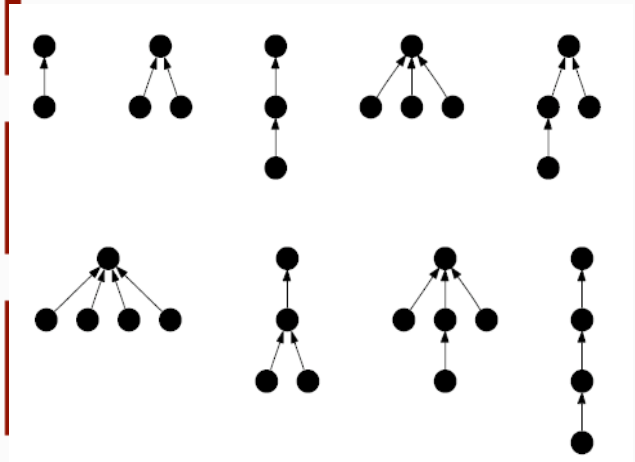


Carnegie Mellon

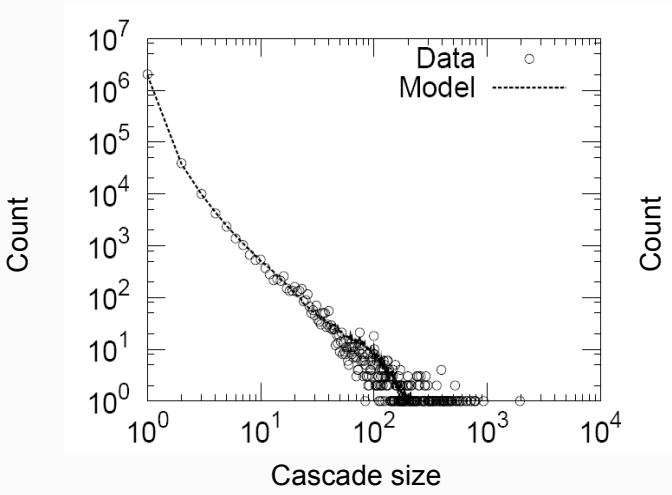
Experimental Results

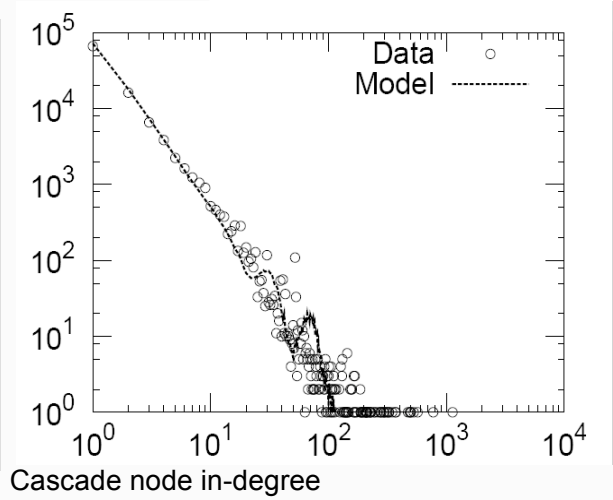
Generative model produces realistic cascades

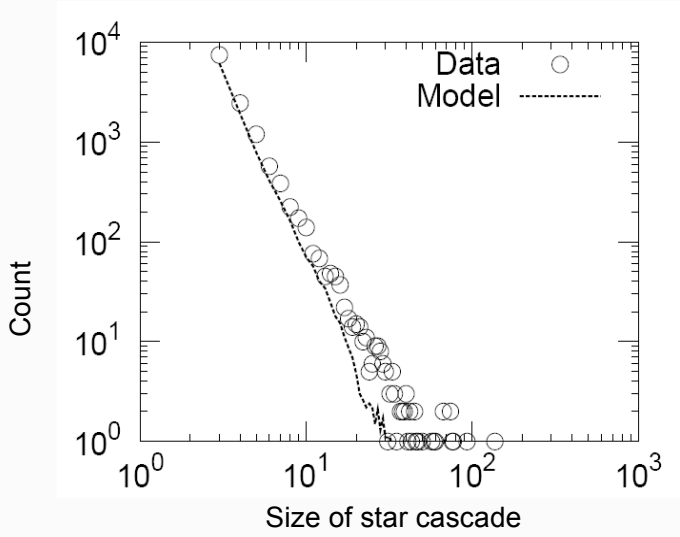
$\beta = 0.025$

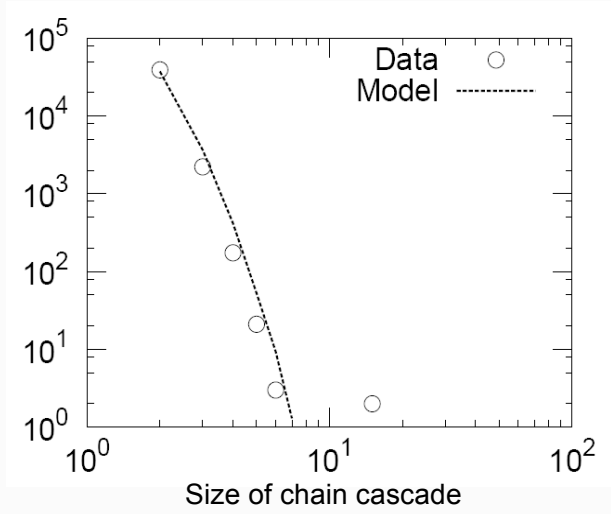


Most frequent cascades









Carnegie Mellon

Conclusions

- Temporal observations
  - Post popularity-dropoff follows power law (exponent= -1.5)
- Topological observations
  - Power-laws in degree distribution, cascade sizes
  - “Stars” are more common than “chains”
- Cascade generating model
  - Based on epidemiology
  - Matches frequent cascades, size power laws

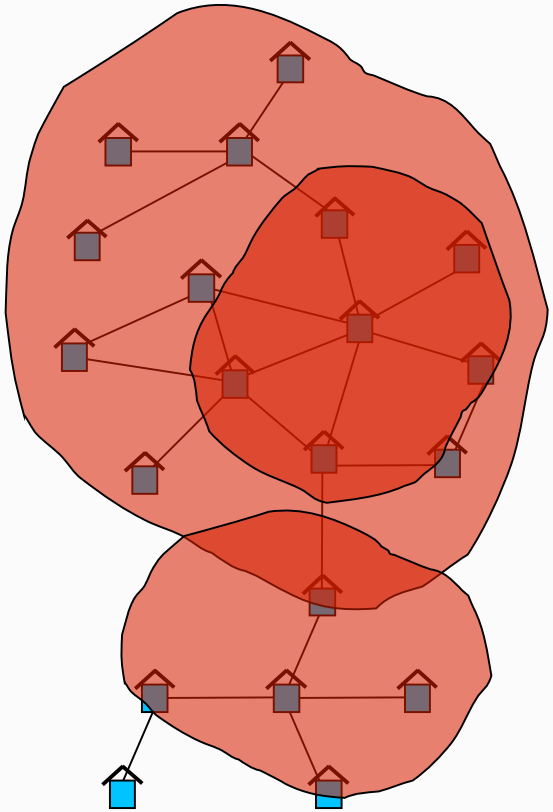
Part 3: Case Studies

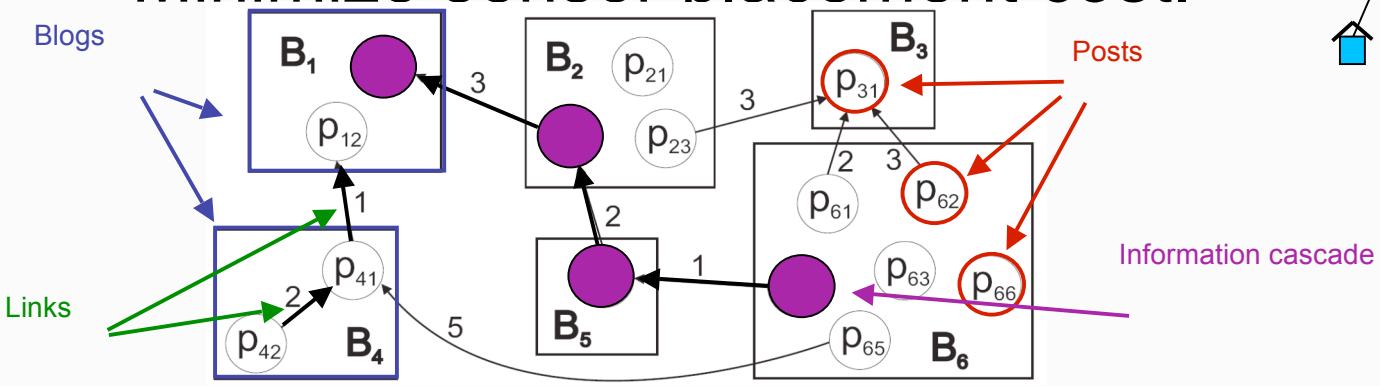
- Q4: How do ideas diffuse through a network?
  - Cascades
  - Epidemiological modeling of cascades
  - Outbreak detection
- Q5: How can we extract communities?
  - Using PCA on structure
  - Factorization
- Q6: What sort of anomaly detection can we perform?
  - Fraud detection on E-bay
  - Spam detection

Carnegie Mellon

Outbreak detection

- Problems of finding sources of contamination in water networks and finding "hot" stories on blogs are isomorphic.
  - Minimize time to detection, population affected
  - Maximize probability of detection.
  - Minimize sensor placement cost.





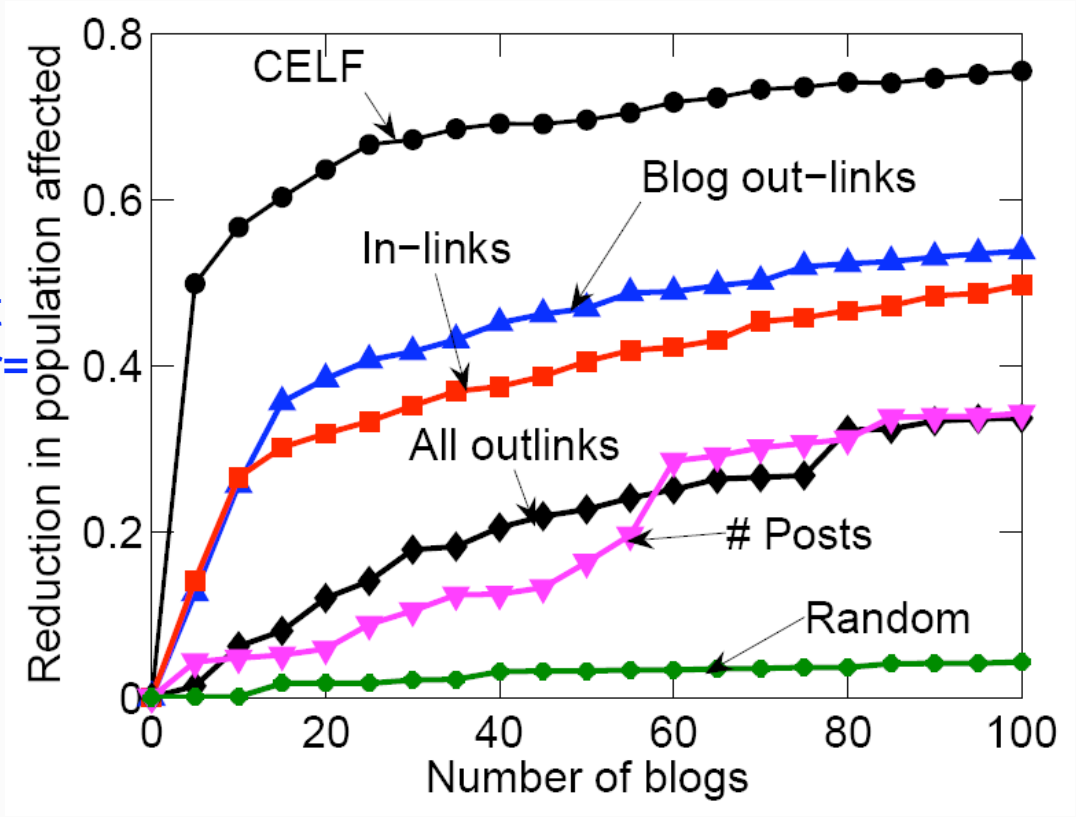
CELF: Main idea

- Given a graph $G(V,E)$
- and a budget of B sensors
- and data on how contaminations spread over the network:
  - for each contamination i we know the time $T(i, u)$ when it contaminated node u
- Minimize time to detect outbreak
- CELF algorithm uses submodularity and lazy evaluation

J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance. "Cost-effective Outbreak Detection in Networks" KDD 2007

Blogs: Comparison to heuristics

Benefit (higher= better)



"Best 10 blogs to read"

http://www.cs.cmu.edu/~jure/blogs/blogs-uc-pa.html

NP - number of posts, IL- in-links, OLO- blog out links, OLA- all out links

k	PA score	Blog	NP	IL	OLO	OLA
1	0.1283	http://instapundit.com	4593	4636	1890	5255
2	0.1822	http://donsurber.blogspot.com	1534	1206	679	3495
3	0.2224	http://sciencepolitics.blogspot.com	924	576	888	2701
4	0.2592	http://www.watcherofweasels.com	261	941	1733	3630
5	0.2923	http://michellemalkin.com	1839	12642	1179	6323
6	0.3152	http://blogometer.nationaljournal.com	189	2313	3669	9272
7	0.3353	http://themodulator.org	475	717	1844	4944
8	0.3508	http://www.bloggersblog.com	895	247	1244	10201
9	0.3654	http://www.boingboing.net	5776	6337	1024	6183
10	0.3778	http://atrios.blogspot.com	4682	3205	795	3102

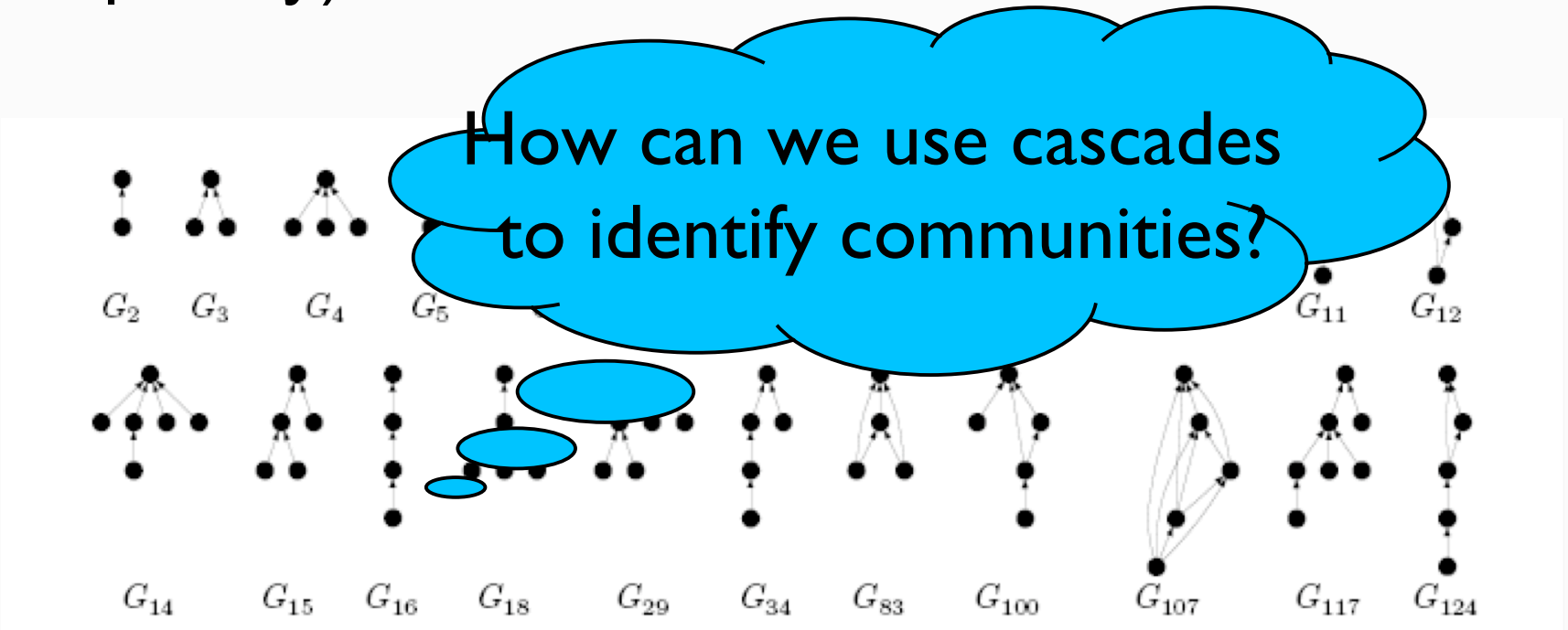
Part 3: Case Studies

- Q4: How do ideas diffuse through a network?
  - Cascades
  - Epidemiological modeling of cascades
  - Outbreak detection
- Q5: How can we extract communities?
  - Using PCA on structure
  - Factorization
- Q6: What sort of anomaly detection can we perform?
  - Fraud detection on E-bay
  - Spam detection

Carnegie Mellon

Blogs and structure

- Cascades take on different shapes (sorted by frequency):

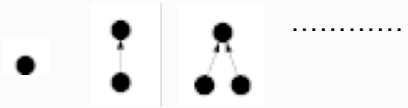


Carnegie Mellon

PCA on cascade types

- Perform PCA on sparse matrix.
- Use log(count+1)
- Project onto 2 PC…

~9,000 cascade types



~44,000 blogs

slashdot	4.6	2.1	.09			
boingboing	3.2	1.1		3.4	.07	
. . .	4.2					
. . .	5.1					
. . .	2.1		1.1			
. . .	.67			.07		
. . .	.01					

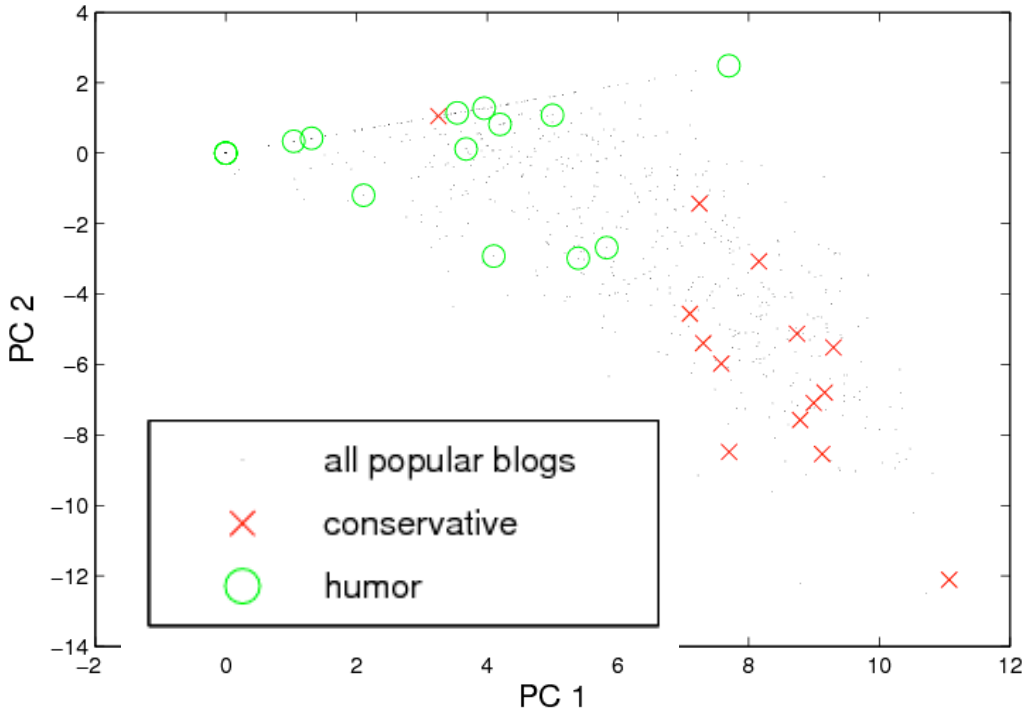
Carnegie Mellon

PCA on cascade types

- Observation: Content of blogs and cascade behavior are often related.

- Distinct clusters for "conservative" and "humorous" blogs (hand-labeling).

M. McGlohon, J. Leskovec, C. Faloutsos, M. Hurst, N. Glance. Finding Patterns in Blog Shapes and Blog Evolution. ICWSM 2007.



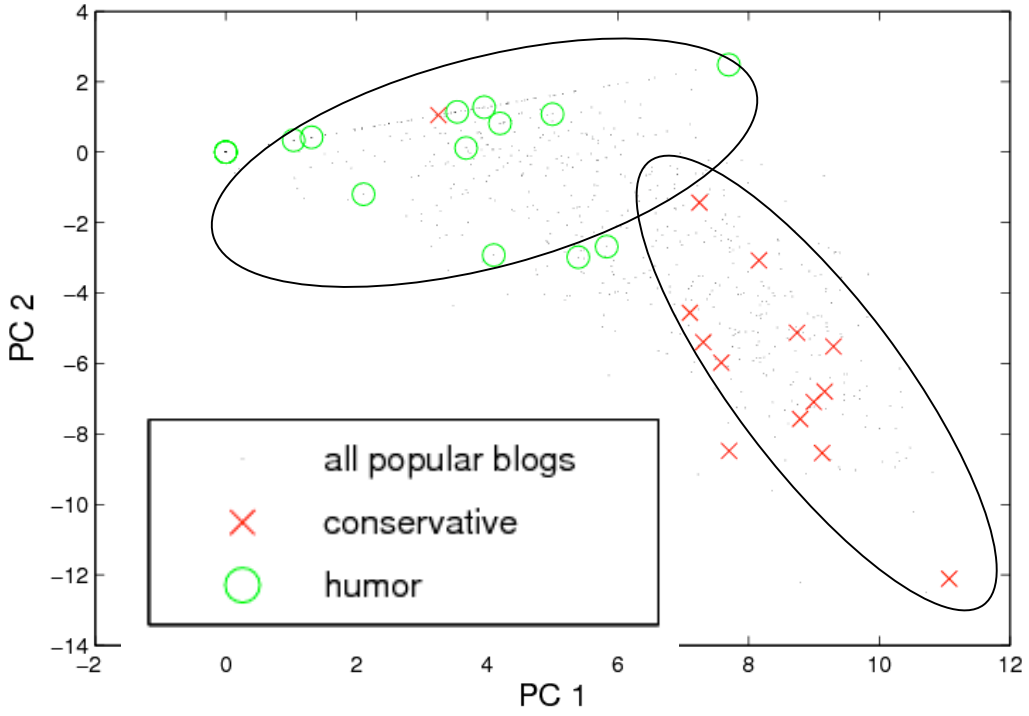
Carnegie Mellon

PCA on cascade types

- Observation: Content of blogs and cascade behavior are often related.

- Distinct clusters for "conservative" and "humorous" blogs (hand-labeling).

M. McGlohon, J. Leskovec, C. Faloutsos, M. Hurst, N. Glance. Finding Patterns in Blog Shapes and Blog Evolution. ICWSM 2007.



Part 3: Case Studies

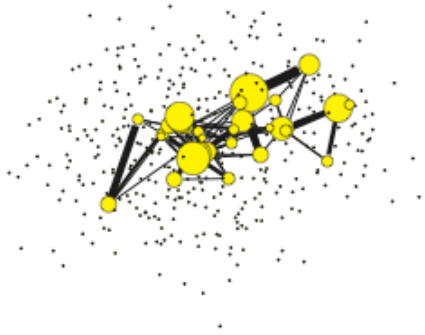
- Q4: How do ideas diffuse through a network?
  - Cascades
  - Epidemiological modeling of cascades
  - Outbreak detection
- Q5: How can we extract communities?
  - Using PCA on structure
  - Factorization
- Q6: What sort of anomaly detection can we perform?
  - Fraud detection on E-bay
  - Spam detection

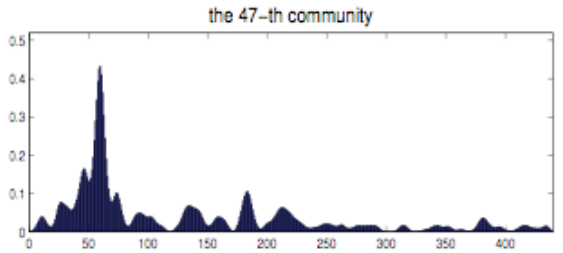
Community Factorization

- Yun Chi, Shenghuo Zhu, Xiaodan Song, Junichi Tatemura, Belle L. Tseng. Structural and temporal analysis of the blogosphere through community factorization. KDD 07
- Main idea: Use tensor factorization to identify subgraphs over time.

Community Factorization Results

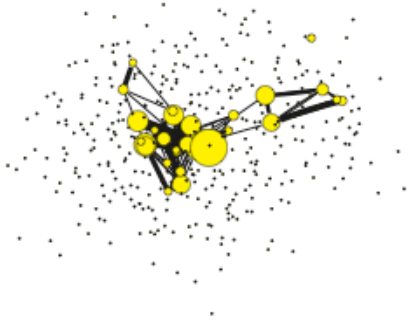
- Hurricane Katrina community

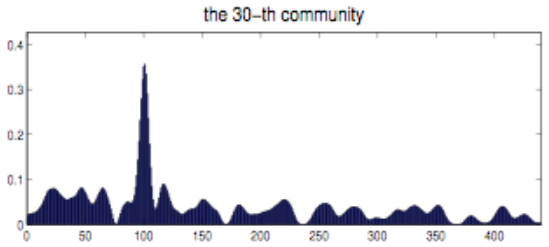




louisiana fema orleans cohen storm malkin hurricane katrina buses volunteers governor mayor guard corps rescue memorial flood michelle wise voices emergency volunteer ties welfare air relief roundup disaster borders homeland rumor loan boards dept supplies wiki flip shelter cross journalist recovery authorities tribune

- Blog info community





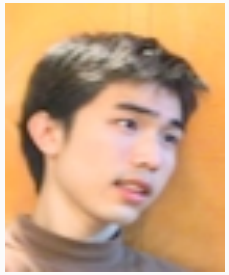
rank linking ranked analysis technorati blind blogosphere log studies ranking relative ranks fake figure conclusions blogs unique tail structured total literature curve tracked methods approaches tends spam welfare partly misleading blogspot vs volume scale weblog statistics collect collected chart profiles

Part 3: Case Studies

- Q4: How do ideas diffuse through a network?
  - Cascades
  - Epidemiological modeling of cascades
  - Outbreak detection
- Q5: How can we extract communities?
  - Using PCA on structure
  - Factorization
- Q6: What sort of anomaly detection can we perform?
  - Fraud detection on E-bay
  - Spam detection

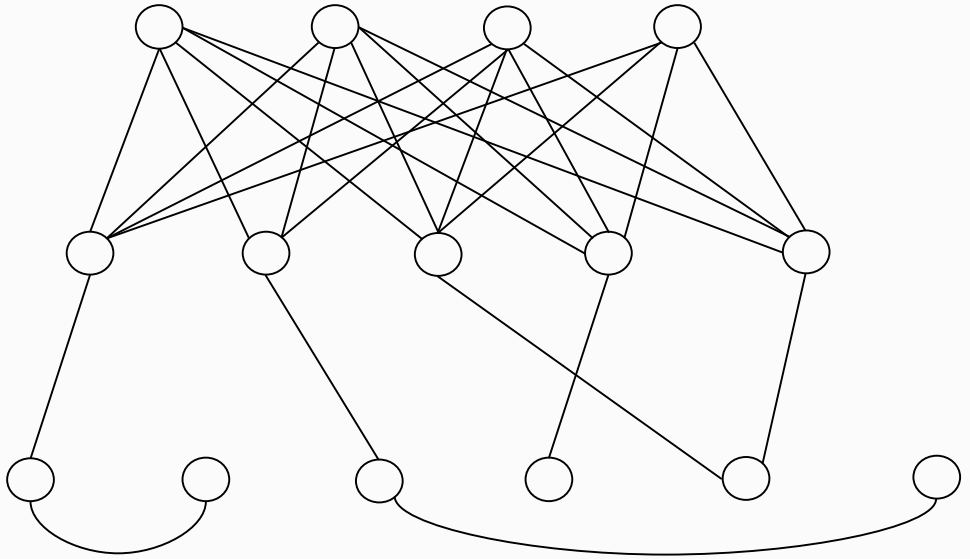
Carnegie Mellon

E-bay Fraud detection



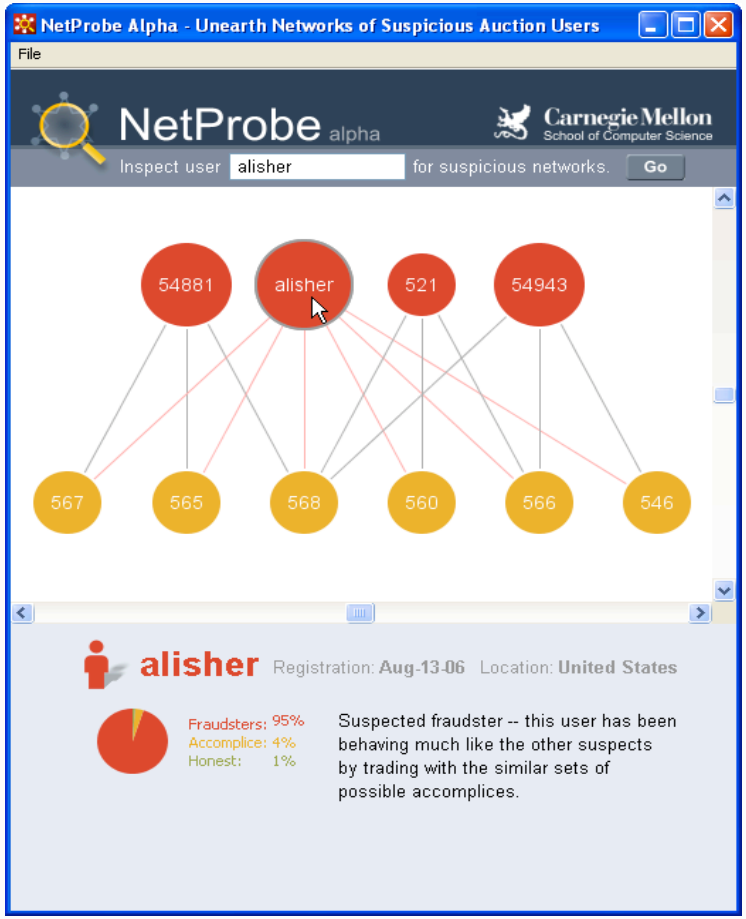
Shashank Pandit, Duen Horng Chau, Samuel Wang, and Christos Faloutsos. NetProbe: A Fast and Scalable System for Fraud Detection in Online Auction Networks WWW 07.

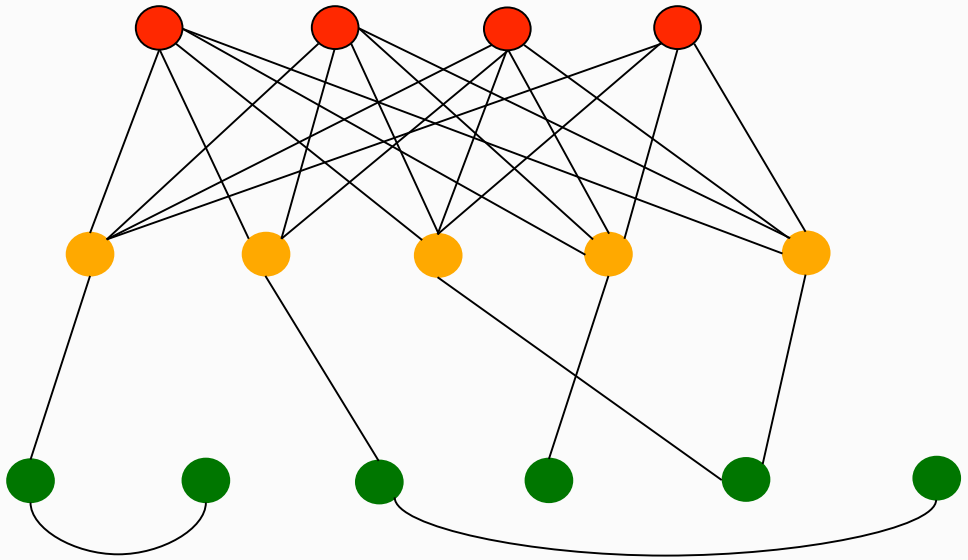
Detects
'non-delivery' fraud:
seller takes $$
and disappears



Carnegie Mellon

E-bay Fraud detection - NetProbe





Idea: 'Accomplices', and Belief Propagation

- 3 types of nodes: honest, fraud, accomplices
- 'Accomplices' never do fraud
  - give high ratings to fraudsters-to-be

Belief propagation intuition:

- If I am honest, my neighbors are either honest or 'accomplices'
- If I'm an accomplice, my neighbors are either honest or fraud

Carnegie Mellon

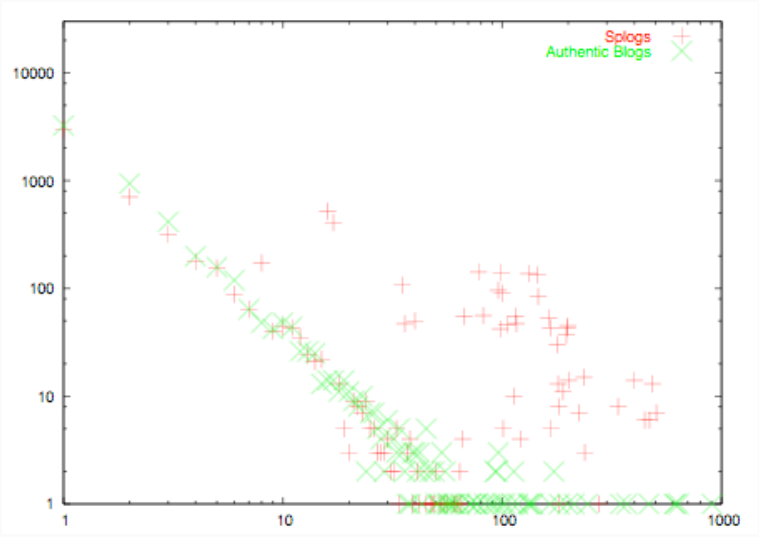
Part 3: Case Studies

- Q4: How do ideas diffuse through a network?
  - Cascades
  - Epidemiological modeling of cascades
  - Outbreak detection
- Q5: How can we extract communities?
  - Using PCA on structure
  - Factorization
- Q6: What sort of anomaly detection can we perform?
  - Fraud detection on E-bay
  - Spam detection

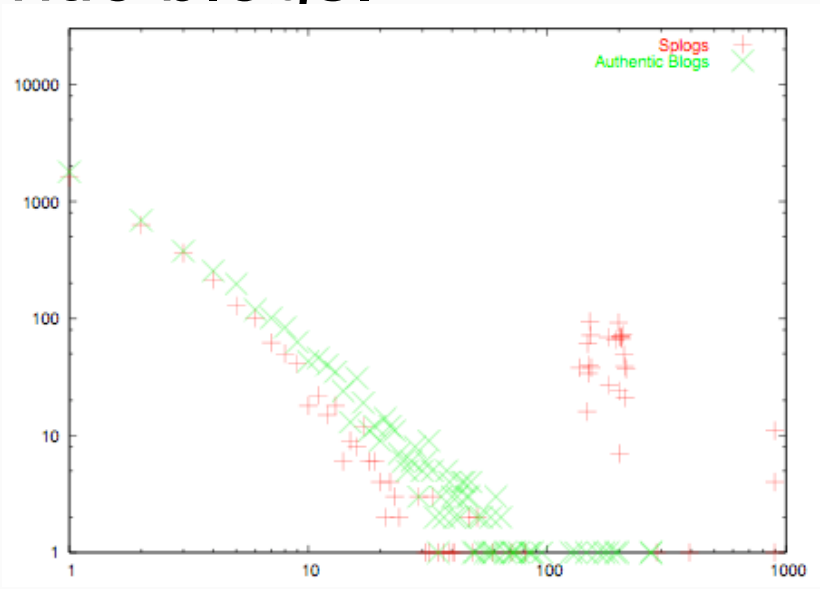
Carnegie Mellon

Spam detection

- Kolari, Java, Finin, 2006:
- Studying link structure can help detect spam in blogs.
- Splogs may deviate from power law degree distribution found in authentic blogs.



indegree



outdegree

Carnegie Mellon

Conclusion

- Presented patterns found in real graphs (power-law degrees, giant connected component, densification, shrinking diameter)
- Demonstrated tools to solve problems (matrix tools, tensors, self-similarity)
- Showed some examples of using these tools for applications to social media (viral marketing, community detection, anomaly detection).

Carnegie Mellon

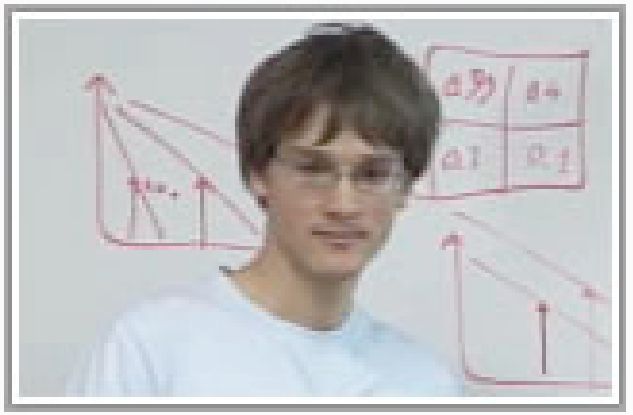
Thanks

- Jimeng Sun (IBM)

- Deepay Chakrabarti (Yahoo!)

- Tamara Kolda (Sandia)

- Jure Leskovec (CMU)



- Epidemiology and viral marketing
  - Adar, E. & Adamic, L. A. (2005), Tracking Information Epidemics in Blogspace, *in* 'WI '05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence', IEEE Computer Society, Washington, DC, USA, pp. 207--214.
  - Bailey, N. (1975), *The Mathematical Theory of Infectious Diseases and its Applications*, Griffin, London.
  - Equiluz, V. M. & Klemm, K. (2002), 'Epidemic threshold in structured scale-free networks', *arXiv:cond-mat/02055439*.
  - Gruhl, D.; Guha, R.; Liben-Nowell, D. & Tomkins, A. (2004), Information Diffusion Through Blogspace, *in* 'WWW '04'.
  - Hethcote, H. W. (2000), 'The Mathematics of Infectious Diseases', *SIAM Rev.* **42**(4), 599--653.

- Kempe, D.; Kleinberg, J. & Tardos, E. (2003), Maximizing the Spread of Influence through a Social Network, *in* 'KDD '03'.
- Leskovec, J.; McGlohon, M.; Faloutsos, C.; Glance, N. & Hurst, M. (2007), Cascading Behavior in Large Blog Graphs: Patterns and a Model, *in* 'Society of Applied and Industrial Mathematics: Data Mining (SDM07)'.
- Leskovec, J.; Singh, A. & Kleinberg, J. (2006), Patterns of Influence in a Recommendation Network, *in* 'Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)'.
- Leskovec, J.; Adamic, L. A. & Huberman, B. A. (2006), The dynamics of viral marketing, *in* 'EC '06: Proceedings of the 7th ACM Conference on Electronic Commerce', ACM Press, New York, NY, USA, pp. 228--237.
- Leskovec, J; Krause, A.; Guestrin, C.; Faloutsos, C.; VanBriesen, J.; Glance, N. Cost-effective outbreak detection in networks. *In* SIGKDD 2007

- Newman, M. E. J. (2005), 'Threshold effects for two pathogens spreading on a network', *Physical Review Letters* **95**, 108701.
- Newman, M. E. J. (2002), 'The spread of epidemic disease on networks', *Physical Review Letters* **66**, 016128.
- Richardson, M. & Domingos, P. (2002),'Mining Knowledge-Sharing Sites for Viral Marketing'.
- Wang, Y.; Chakrabarti, D.; Wang, C. & Faloutsos, C. (2003), Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint., *in* 'SRDS', pp. 25-34.

- Community detection
  - Chi, Y.; Zhu, S.; Song, X.; Tatemura, J. & Tseng, B. L. (2007), Structural and temporal analysis of the blogosphere through community factorization, *in* 'KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, New York, NY, USA, pp. 163--172.
  - McGlohon, M.; Leskovec, J.; Faloutsos, C.; Hurst, M. & Glance, N. (2007), Finding Patterns in Blog Shapes and Blog Evolution, *in* 'International Conference on Weblogs and Social Media'.

- Spam/Anomaly detection
  - Kolari, P.; Java, A. & Finin, T. (2006), Characterizing the Splogosphere, *in* 'Proceedings of the 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 15th World Wide Web Conference', University of Maryland, Baltimore County, .
  - Pandit, S.; Chau, D. H.; Wang, S. & Faloutsos, C. (2007), Netprobe: a fast and scalable system for fraud detection in online auction networks, *in* 'WWW '07: Proceedings of the 16th international conference on World Wide Web', ACM, New York, NY, USA, pp. 201--210.

Carnegie Mellon

Questions?

- Mary McGlohon
  mmcgloho@cs.cmu.edu
  www.cs.cmu.edu/~mmcgloho

- Christos Faloutsos
  christos@cs.cmu.edu
  www.cs.cmu.edu/~christos