Jure Leskovec
Machine Learning Department
Carnegie Mellon University

How to detect epidemics and influential blogs?

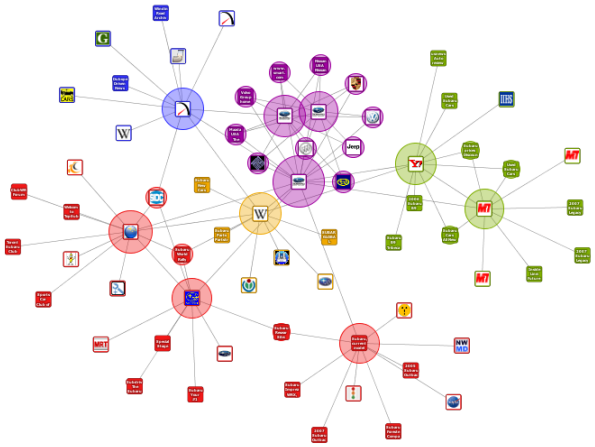
Currently: Carnegie Mellon Soon: 

Networks: Rich data

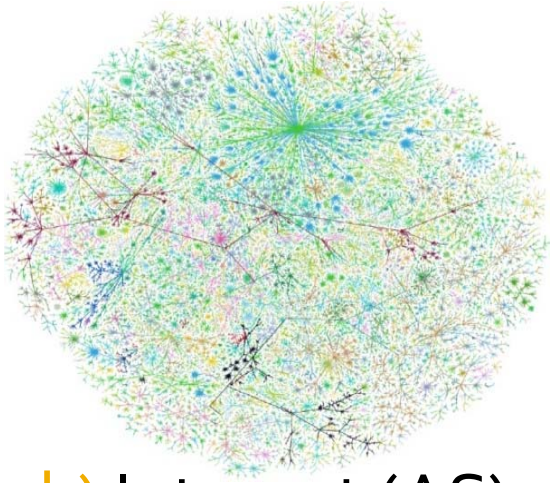
- **Today: L**arge on-line systems have detailed records of human activity
  - On-line communities:
    - Facebook (64 million users, billion dollar business)
    - MySpace (300 million users)
  - Communication:
    - Instant Messenger (~1 billion users)
  - News and Social media:
    - Blogging (250 million blogs world-wide, presidential candidates run blogs)
  - On-line worlds:
    - World of Warcraft (internal economy 1 billion USD)
    - Second Life (GDP of 700 million USD in '07)

Opportunities for **impact** in **science** and **industry**

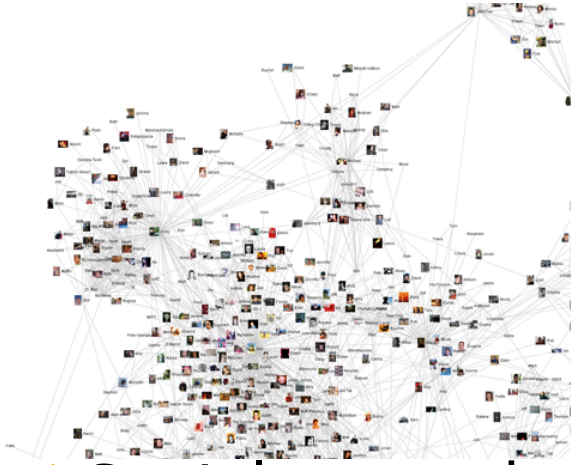
Networks: Rich and massive data



a) World wide web



b) Internet (AS)



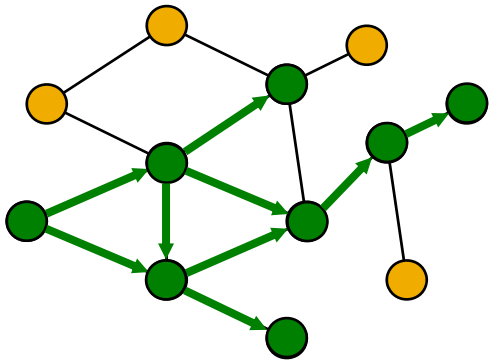
c) Social networks

We need massive network data for the patterns to emerge:

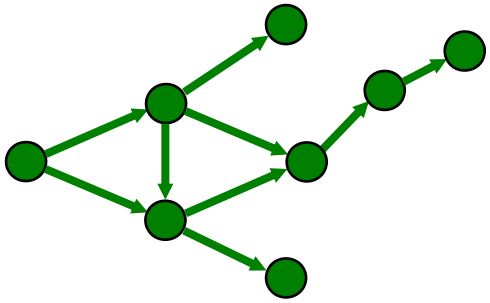
- MSN Messenger network [WWW '08]
  - 240M people, 255B messages, 4.5 TB data
- Blogosphere
  - 60M posts, 120M links

Diffusion and Cascades

- Behavior that cascades from node to node like an epidemic
  - News, opinions, rumors
  - Word-of-mouth in marketing
  - Infectious diseases
- As activations spread through the network they leave a trace – **a cascade**



Network



Cascade
(propagation graph)

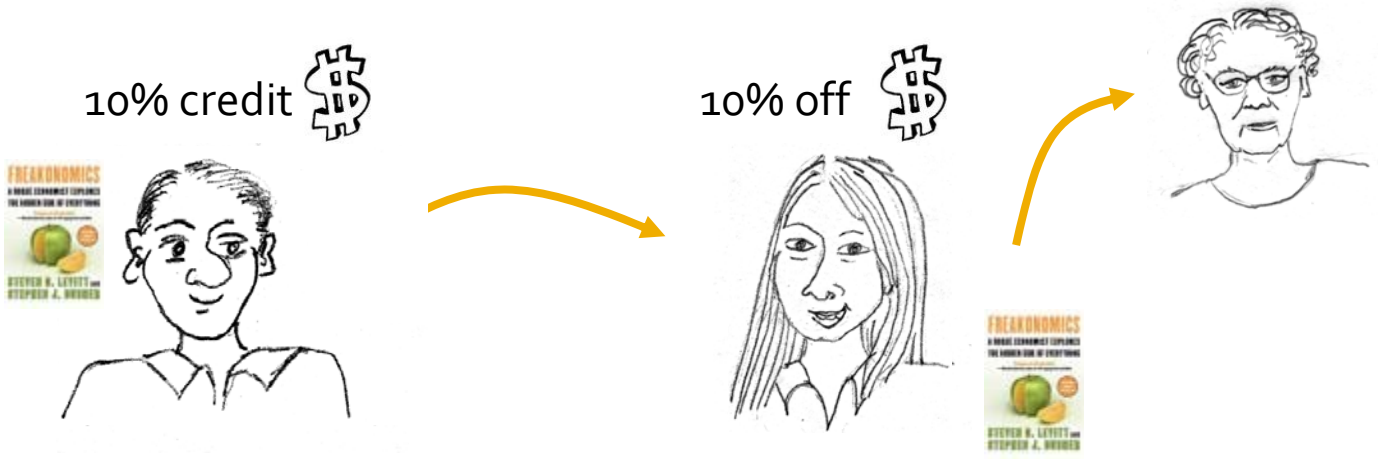
Talk outline

- Where do cascades occur?
  - On the Web we can actually observe and measure a number of cascades
- What do cascades look like?
  - How do information and influence spread?
- How to detect who is influential?
  - Effective and efficient algorithms
  - Saving lives

[w/ Adamic-Huberman, EC '06]

Setting 1: Viral marketing

- People send and receive product recommendations, purchase products

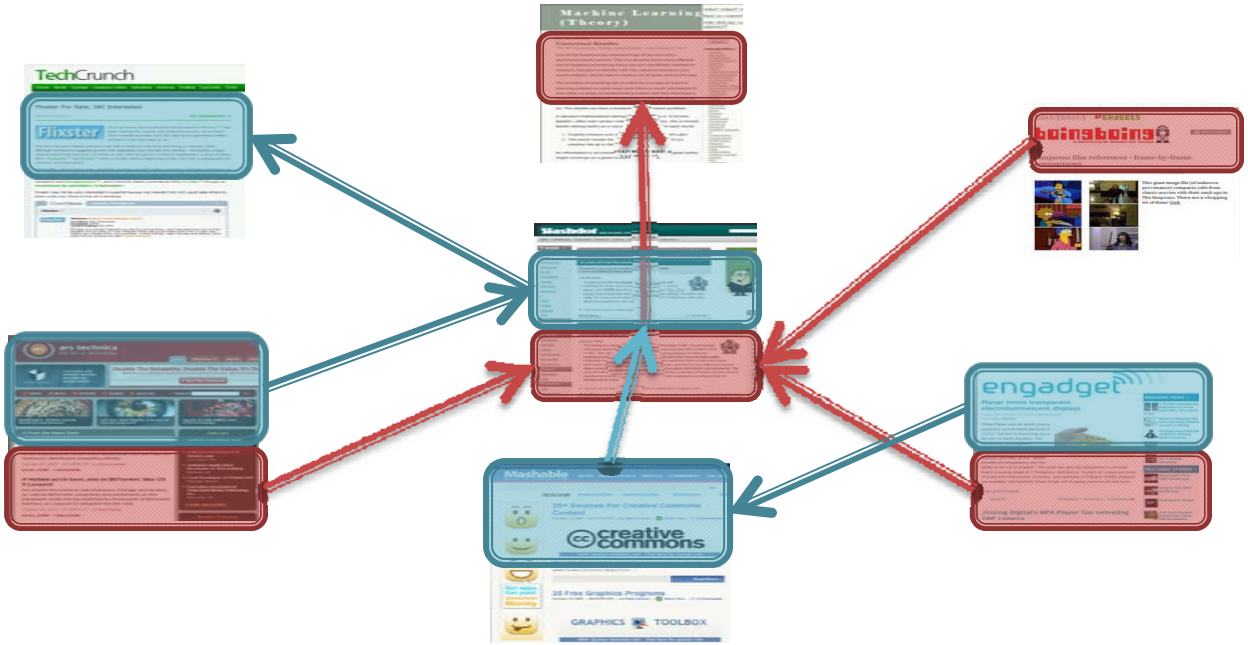


- Data: Large online retailer: 4 million people, 16 million recommendations, 500k products

[w/ Glance-Hurst et al., SDM '07]

Setting 2: Blogosphere

- Bloggers write posts and refer (link) to other posts and the information propagates

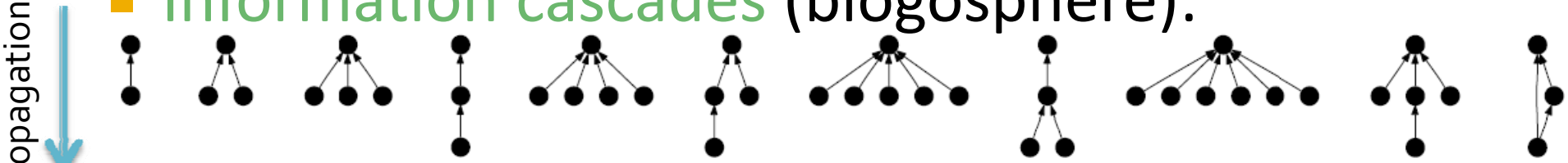


- Data: 10.5 million posts, 16 million links

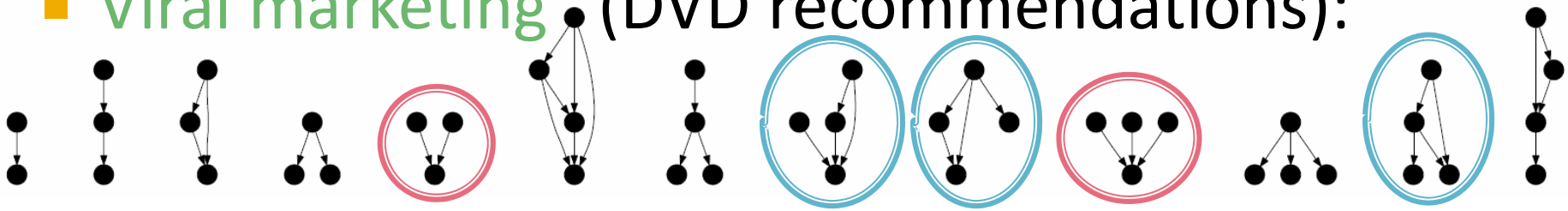
[w/ Kleinberg-Singh, PAKDD '06]

What do cascades look like?

- Are they stars? Chains? Trees?
- Information cascades (blogosphere):



- Viral marketing (DVD recommendations):

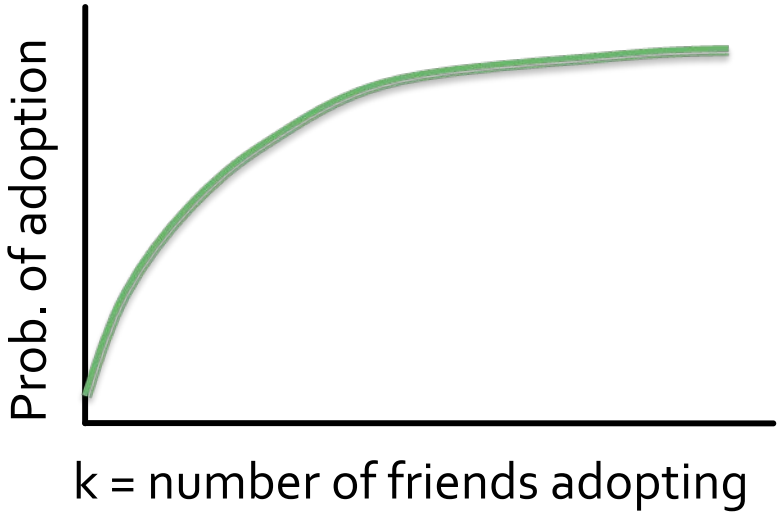


(ordered by frequency)

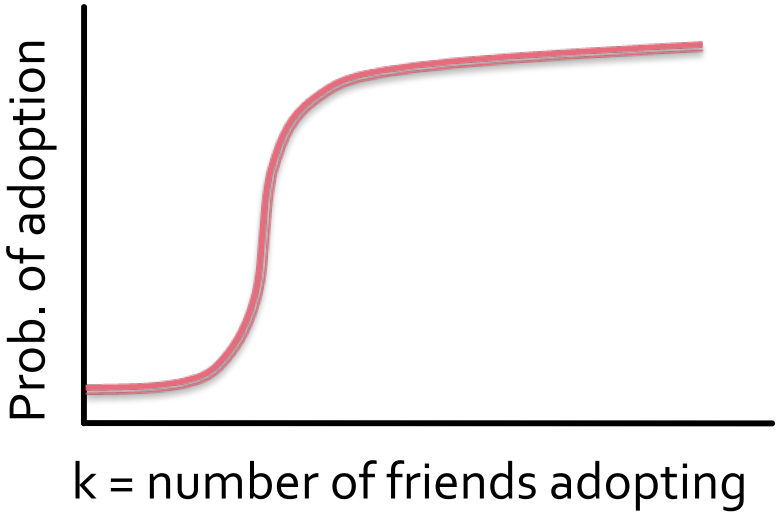
- Viral marketing cascades are more social:
  - Collisions (no summarizers)
  - Richer non-tree structures

Human adoption curves

- Prob. of adoption depends on the number of friends who have adopted [Bass '69, Granovetter '78]
- **What is the shape?**
  - Distinction has consequences for models and algorithms



Diminishing returns?

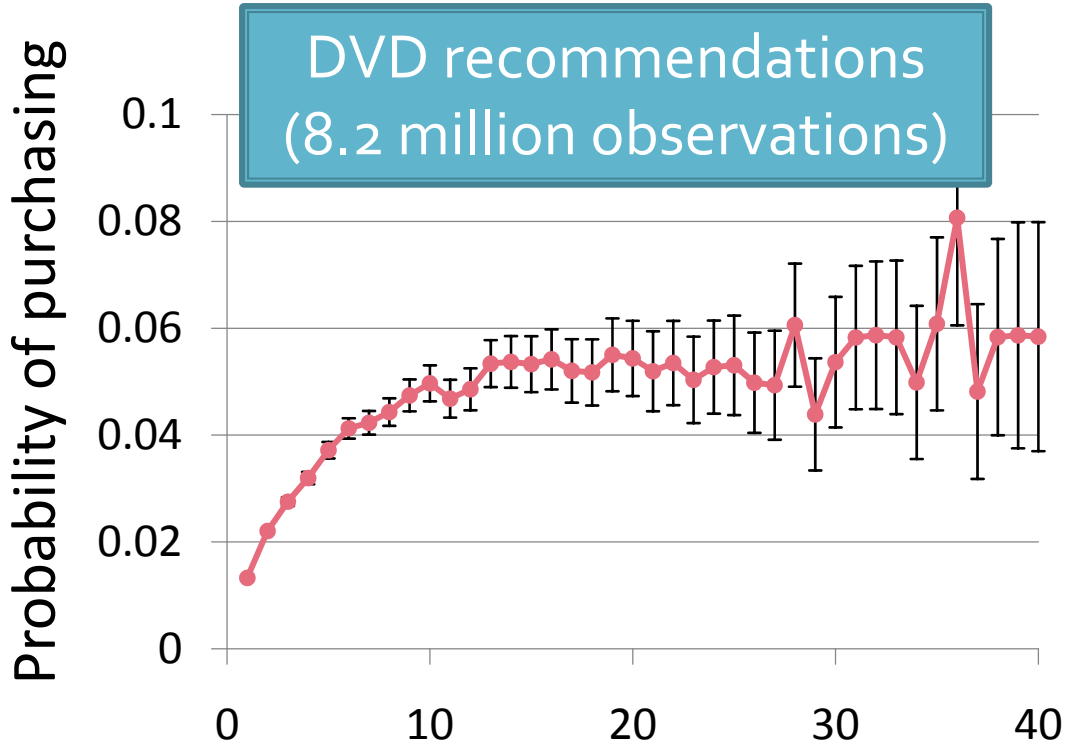


Critical mass?

To find the answer we need lots of data

[w/ Adamic-Huberman, EC '06]

Adoption curve: Validation

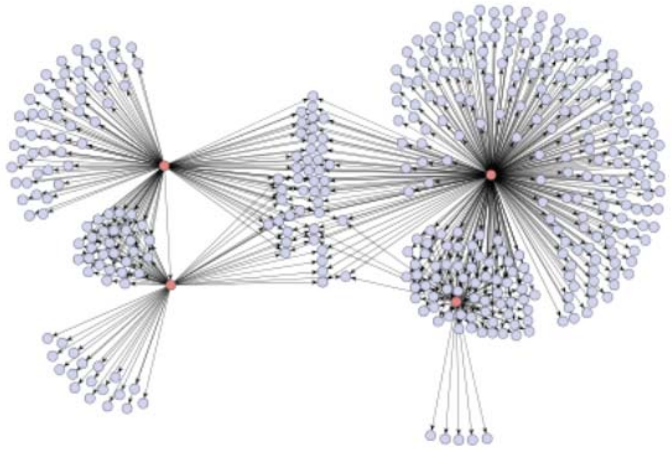


Adoption curve follows the **diminishing returns**.
Can we exploit this?

Later similar findings were made for group membership [Backstrom-Huttenlocher-Kleinberg '06], and probability of communication [Kossinets-Watts '06]

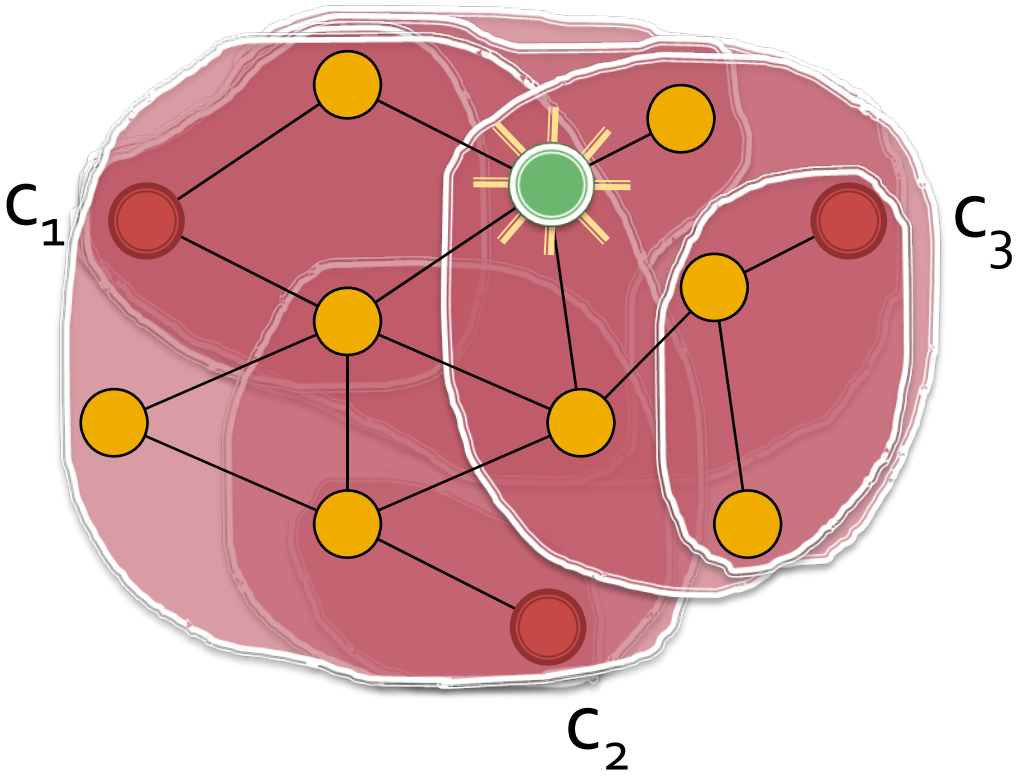
Cascade & outbreak detection

- Blogs – information epidemics
  - Which are the influential/infectious blogs?
- Viral marketing
  - Who are the trendsetters?
  - Influential people?
- Disease spreading
  - Where to place monitoring stations to detect epidemics?



[w/ Krause-Guestrin et al., KDD '07]
(best student paper)

The problem: Detecting cascades

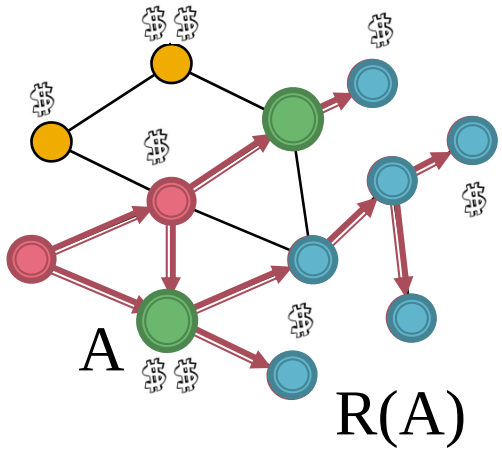


How to quickly detect cascades as they spread?

[w/ Krause-Guestrin et al., KDD '07]
(best student paper)

Two parts to the problem

- Cost:
  - Cost of monitoring is blog dependent (big blogs cost more time to read)
- Reward:
  - Minimize the number of people that that know the story before we do



[w/ Krause-Guestrin et al., KDD '07]
(best student paper)

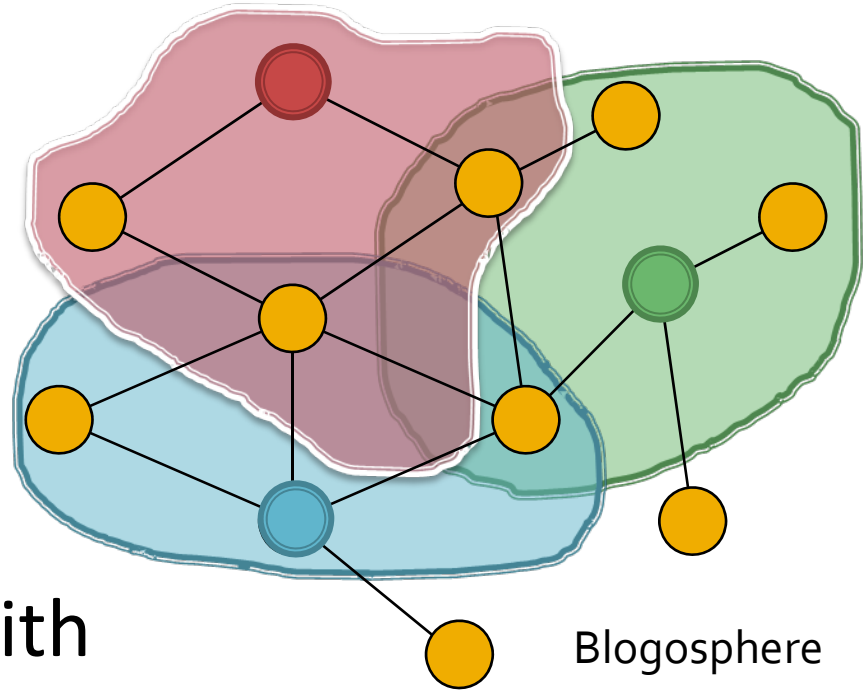
The solution: Covering blogs

= Given a budget (e.g., of 3 blogs)

= Select blogs to cover the most of the blogosphere?

= Bad news: Solving this exactly is **NP-hard**

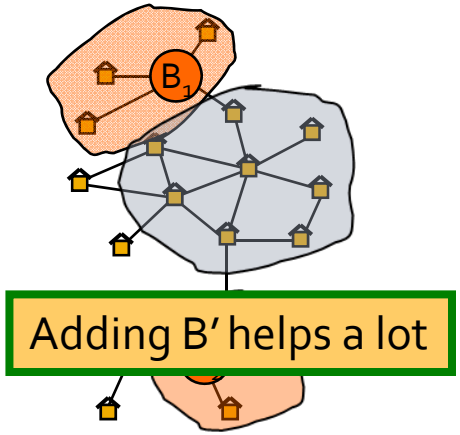
= Good news: **Theorem**: Our algorithm **CELF** can do it in **linear time** and with **factor 3 approximation**



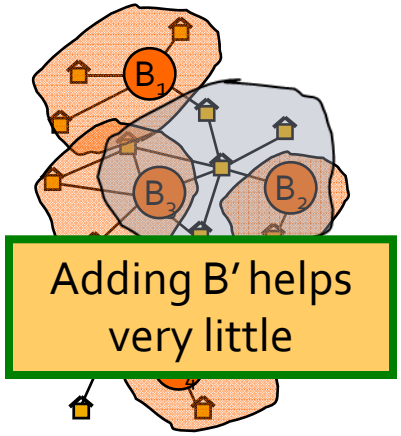
Blogosphere

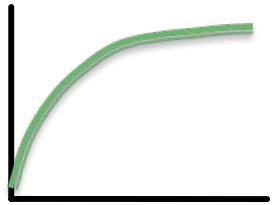
[w/ Krause-Guestrin et al., KDD '07]
(best student paper)

Problem structure: Submodularity



New monitored blog:





Placement A={B_1, B_2}

Placement B={B_1, B_2, B_3, B_4}

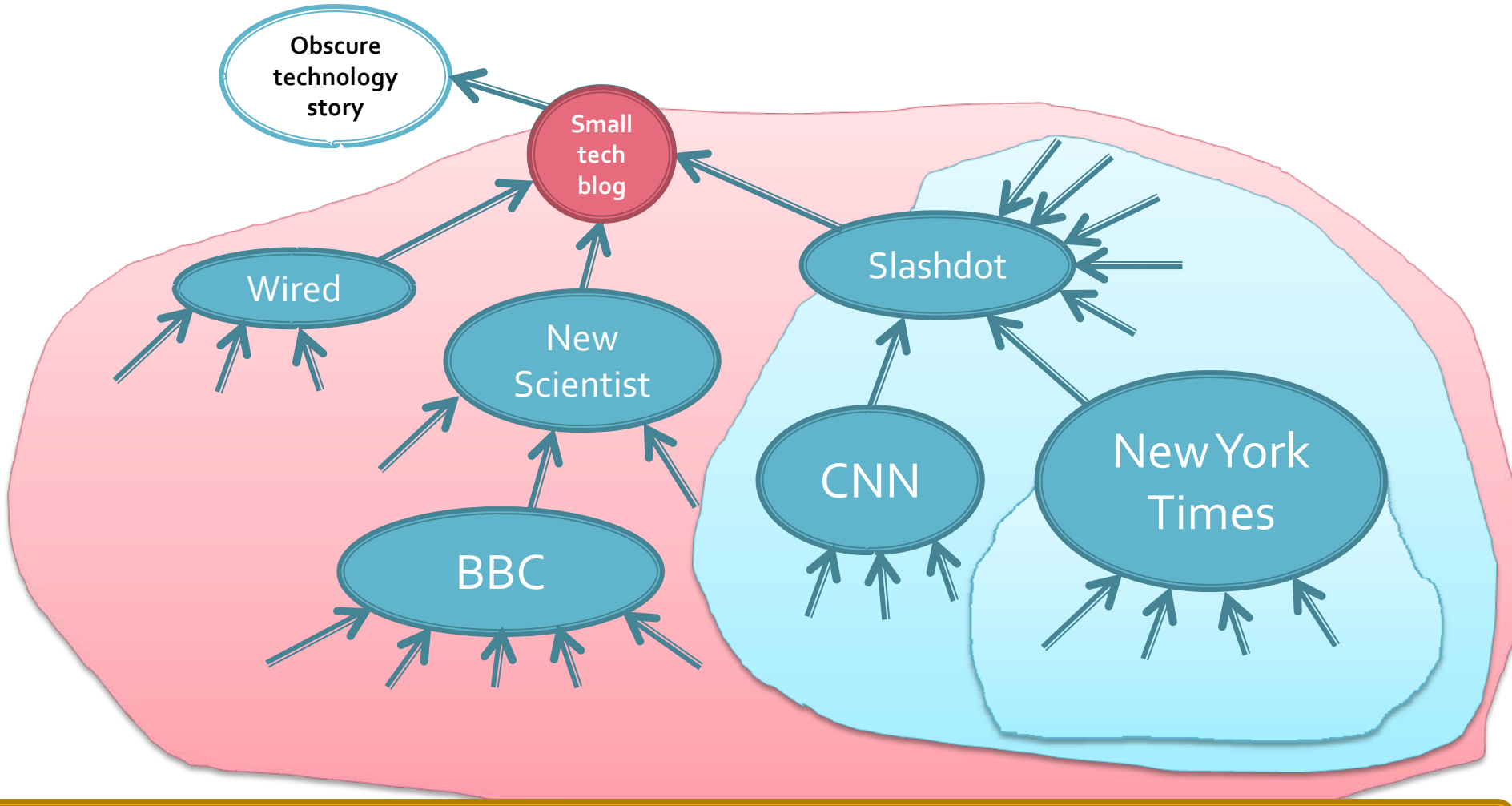
- Gain of adding a node to small set is larger than gain of adding a node to large set
- **Submodularity**: diminishing returns, think of it as "concavity")

Back to the Question…

= I have 10 minutes. Which blogs should I read to be most up to date?

= Who are the most influential bloggers?

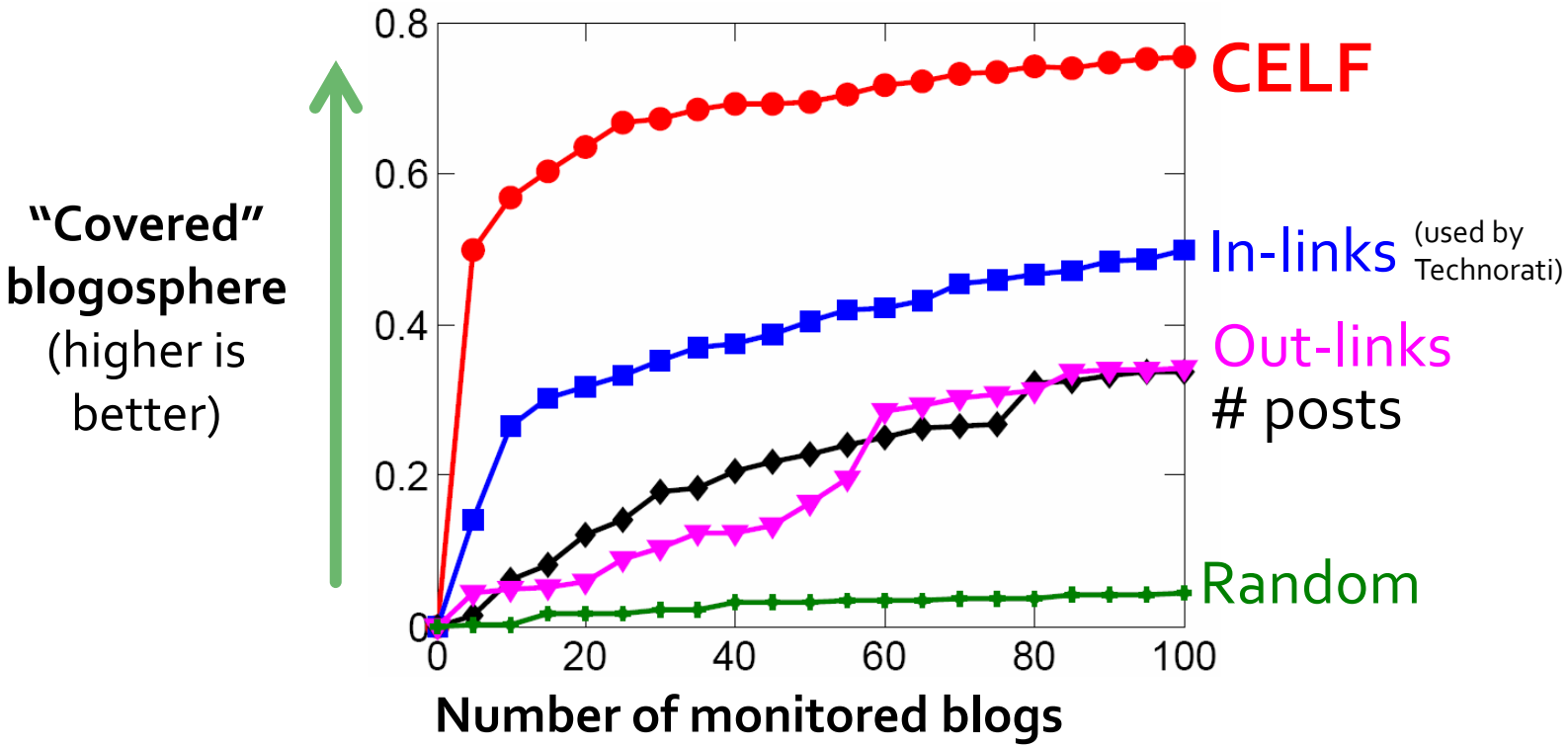
A single story propagates...



Sooner we read the story, more of its influence area we cover

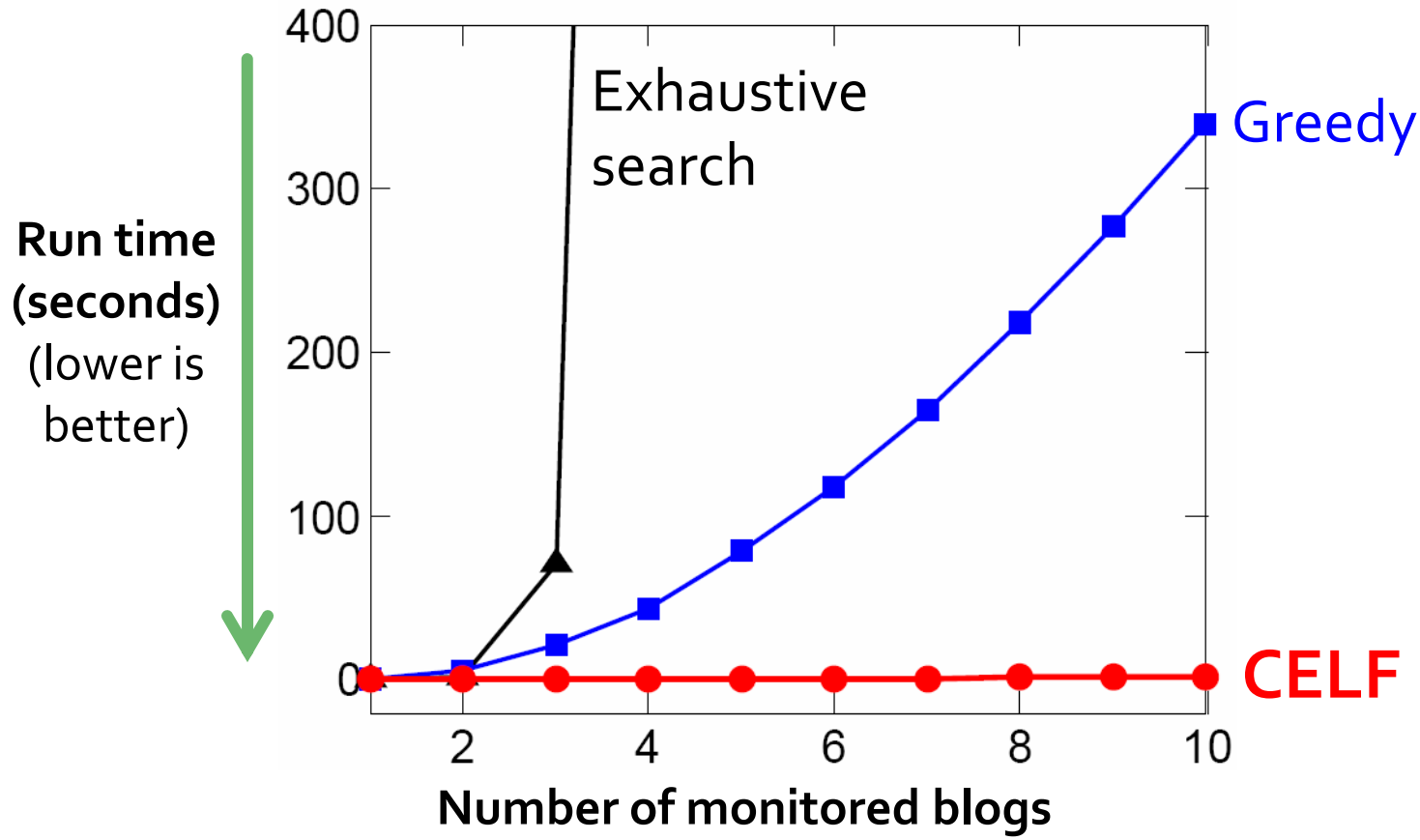
Blogs: Information epidemics

- Which blogs should one read?



For more info see our website: www.blogcascades.org

CELF: Scalability



CELF runs **700x** faster than simple greedy algorithm

So, who is influential? What should I read?

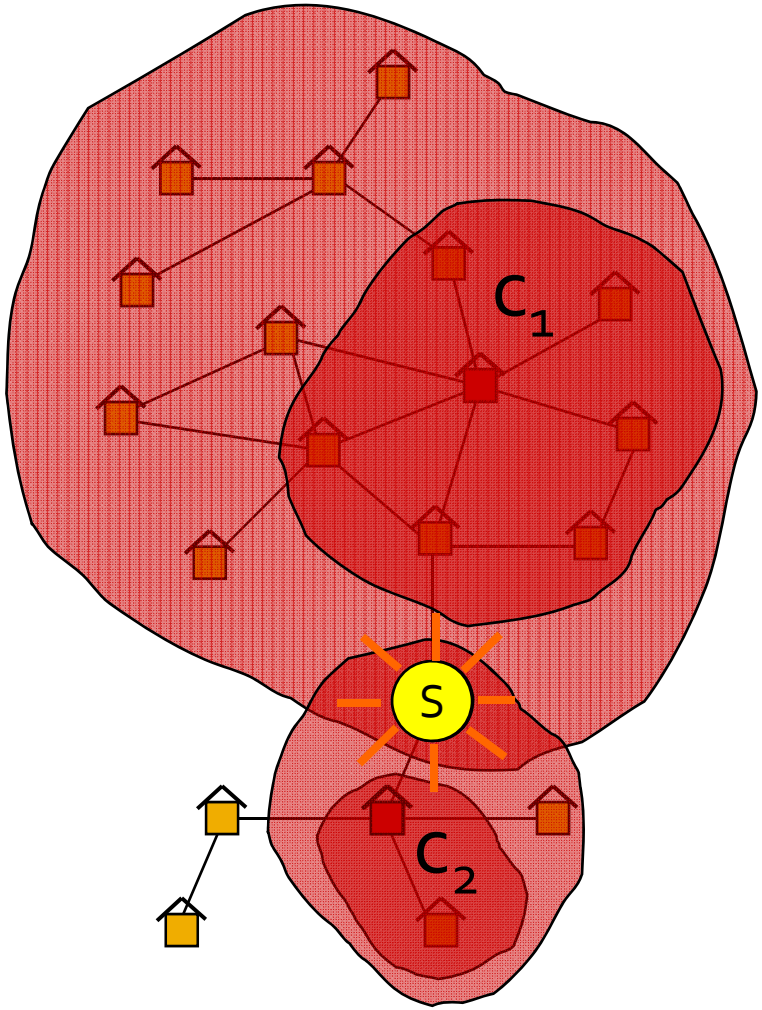
www.blogcascades.org

k	Score	Blog	Posts	InLinks	OutLinks
1	0.13	http://instapundit.com	4593	4636	5255
2	0.18	http://donsurber.blogspot.com	1534	1206	3495
3	0.22	http://sciencepolitics.blogspot.com	924	576	2701
4	0.26	http://www.watcherofweasels.com	261	941	3630
5	0.29	http://michellemalkin.com	1839	12642	6323
6	0.32	http://blogometer.nationaljournal.com	189	2313	9272
7	0.34	http://themodulator.org	475	717	4944
8	0.35	http://www.bloggersblog.com	895	247	10201
9	0.37	http://www.boingboing.net	5776	6337	6183
10	0.38	http://atrios.blogspot.com	4682	3205	3102
11	0.39	http://lawhawk.blogspot.com	1862	463	6597
12	0.40	http://www.gothamist.com	6223	3324	17172
13	0.41	http://mparent7777.livejournal.com	25925	199	47933
14	0.42	http://wheelgun.blogspot.com	1174	128	939
15	0.43	http://gevkaffeegal.typepad.com/the_alliance	302	428	2481

[w/ Krause et al., J. of Water Resource Planning]

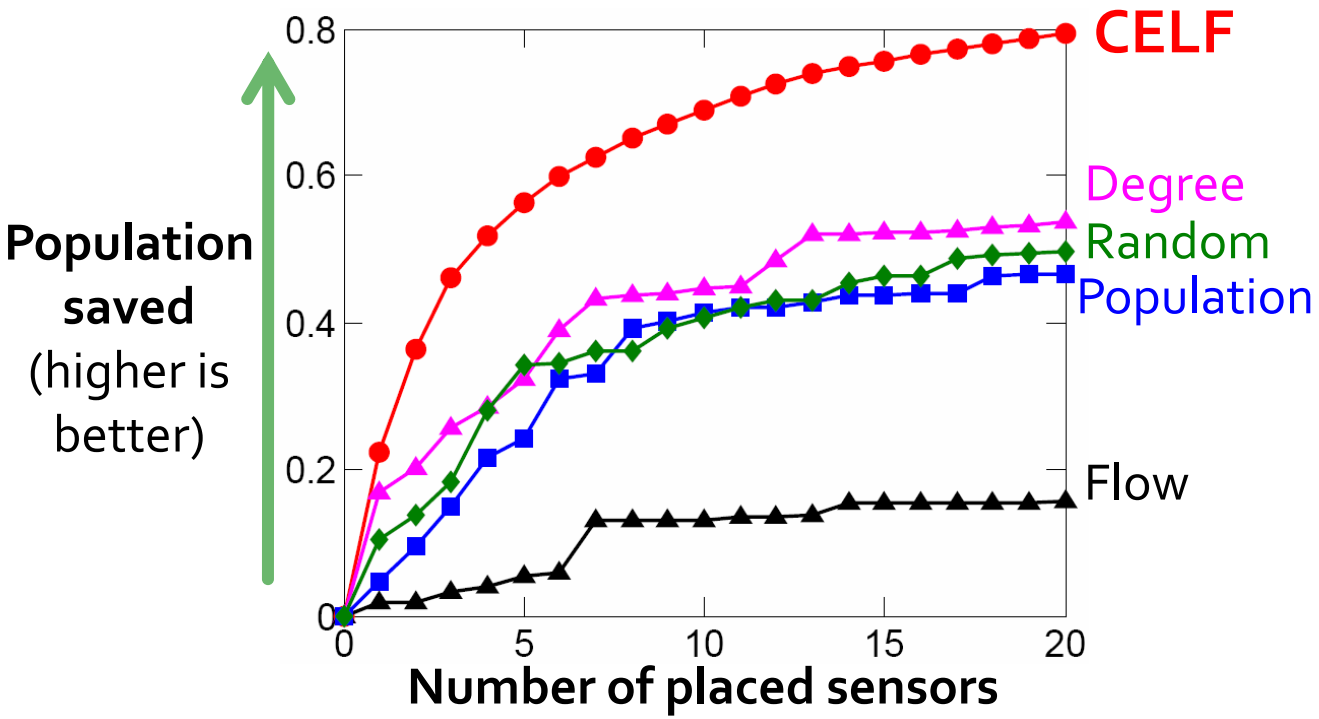
Same problem: Water Network

- Given:
  - a real city water distribution network
  - data on how contaminants spread over time
- Place sensors (to save lives)
- Problem posed by the *US Environmental Protection Agency*



[w/ Ostfeld et al., J. of Water Resource Planning]

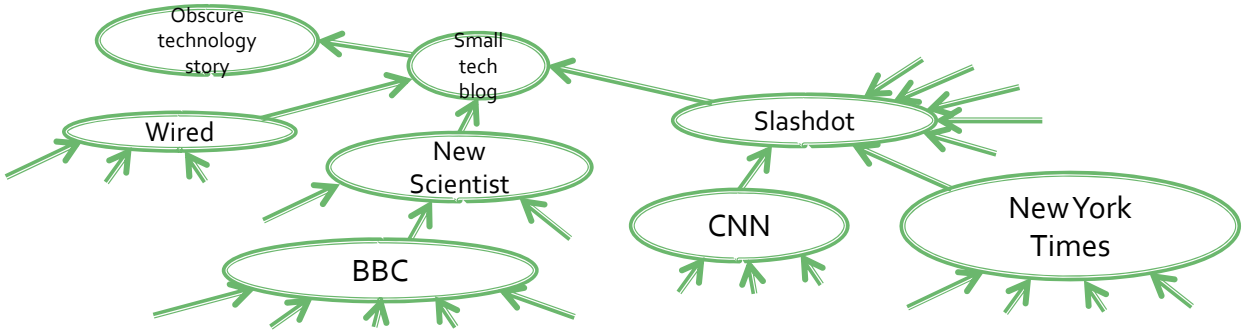
Water network: Results

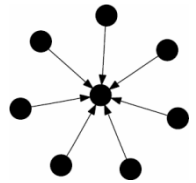


Author	Score
CMU (CELF)	**26**
Sandia	21
U Exter	20
Bentley systems	19
Technion (1)	14
Bordeaux	12
U Cyprus	11
U Guelph	7
U Michigan	4
Michigan Tech U	3
Malcolm	2
Proteo	2
Technion (2)	1

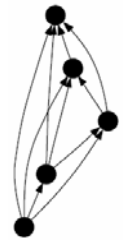
- Our approach performed best at the Battle of Water Sensor Networks competition

Conclusion and connections





- How do news and information spread
  - New ranking and influence measures for blogs
  - Recommendations and incentives
  - Diffusion of topics (news, media)
- Predictive models of information diffusion
  - Social Media Marketing
- How to design better systems incorporating diffusion and incentives



References

- Jure Leskovec, jure@cs.cmu.edu
- http://www.cs.cmu.edu/~jure/

- Jure Leskovec, Lada Adamic, Bernardo Huberman. The Dynamics of Viral Marketing. ACM TWEB 2007.
- Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, Matthew Hurst. Cascading Behavior in Large Blog Graphs. SIAM Data Mining 2007.
- Jure Leskovec, Ajit Singh, Jon Kleinberg. Patterns of Influence in a Recommendation Network. PAKDD 2006.
- Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, Natalie Glance. Cost-effective Outbreak Detection in Networks. ACM KDD, 2007.