Jure Leskovec
Machine Learning Department
Carnegie Mellon University

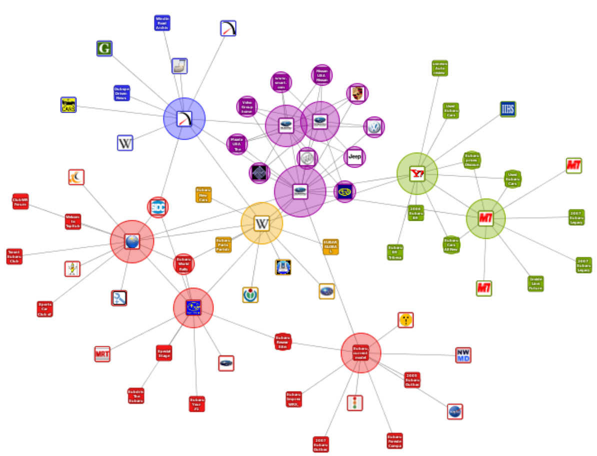
Dynamics of large networks

Web: Rich data

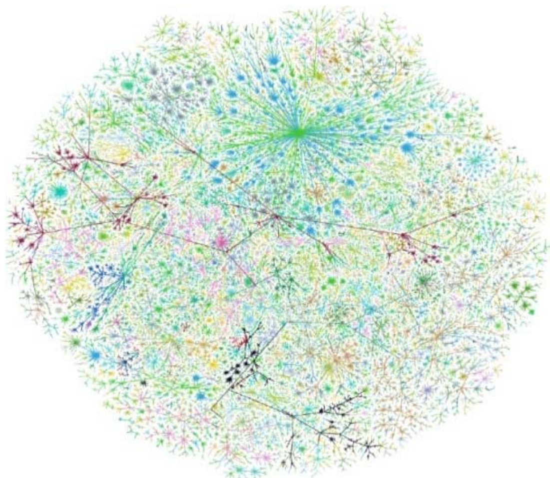
- **Today:** Large on-line systems have detailed records of human activity
  - On-line communities:
    - Facebook (64 million users, billion dollar business)
    - MySpace (300 million users)
  - Communication:
    - I
  - Ne
    - al
  - Or
    - S

Can study phenomena and behaviors at scales that before were never possible

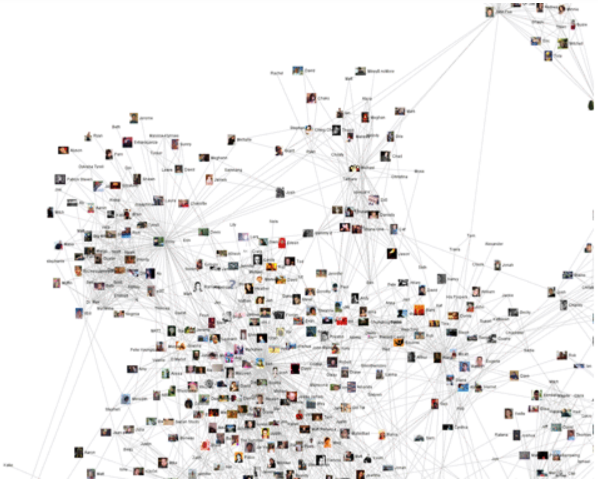
Rich data: Networks



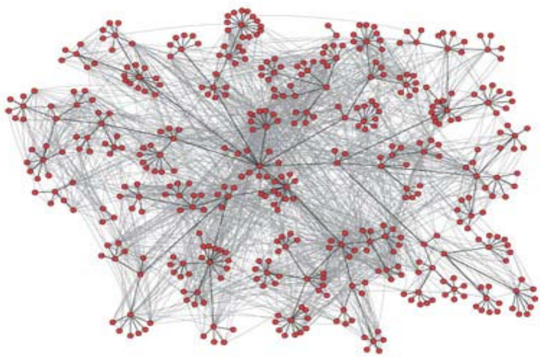
a) World wide web



b) Internet (AS)

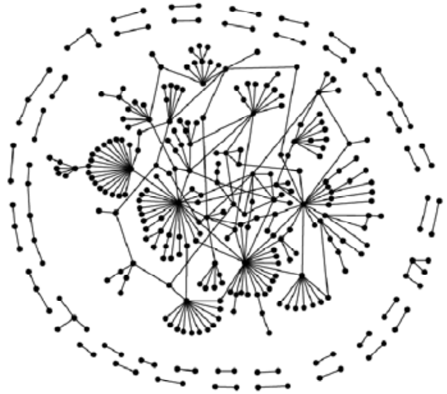


c) Social networks



d) Communication

e) Citations



f) Biological networks

Networks: What do we know?

- We know lots about the network structure:
  - Properties: Scale free [Barabasi ’99], Clustering [Watts-Strogatz ‘98], Navigation [Adamic-Adar ’03, LibenNowell ’05], Bipartit… …twork motifs [Milo et a… …, Conducta… …06], Hubs and autho… …9]
  - Models: P… …99], Small-world [Wa… …[Kleinberg el al. ’99], Heuristically optimized tradeoffs [Fabrikant et al. ‘02], Congestion [Mihail et al. ‘03], Searchability [Kleinberg ‘00], Bowtie [Broder et al. ‘00], Transit-stub [Zegura ‘97], Jellyfish [Tauro et al. ‘01]

We know much less about **processes** and **dynamics** of **networks**

This thesis: Network dynamics

- Network evolution
  - How network structure changes as the network grows and evolves?
- Diffusion and cascading behavior
  - How do rumors and diseases spread over networks?
- Large data
  - Observe phenomena that is “invisible” at smaller scales

This thesis: The structure

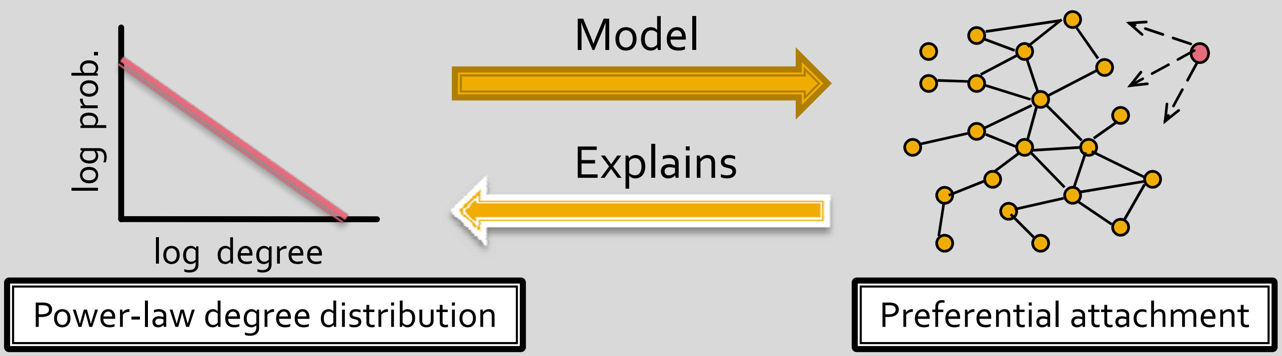
	Network Evolution	Network Cascades	Large Data
Observations	**Q1:** How does network structure evolve over time?	**Q4:** What are patterns of diffusion in networks?	**Q7:** What are the properties of a social network of the whole planet?
Models	**Q2:** How to model individual edge attachment?	**Q5:** How do we model influence propagation?	**Q8:** What is community structure of large networks?
Algorithms (applications)	**Q3:** How to generate realistic looking networks?	**Q6:** How to identify influential nodes and epidemics?	**Q9:** How to predict search result quality from the web graph?

This thesis: The structure

	Network Evolution	Network Cascades	Large Data
Observations	Q1: How does network structure evolve over time?	Q4: What are patterns of diffusion in networks?	Q7: What are the properties of a social network of the whole planet?
Models	Q2: How to model individual edge attachment?	Q5: How do we model influence propagation?	Q8: What is community structure of large networks?
Algorithms (applications)	Q3: How to generate realistic looking networks?	Q6: How to identify influential nodes and epidemics?	Q9: How to predict search result quality from the web graph?

Background: Network models

- Empirical findings on real graphs led to new network models



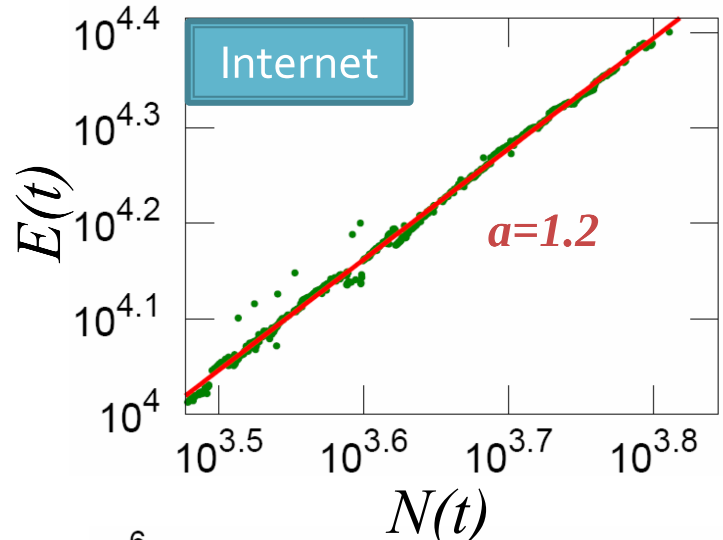
- Such models make assumptions/predictions about other network properties
- What about network evolution?

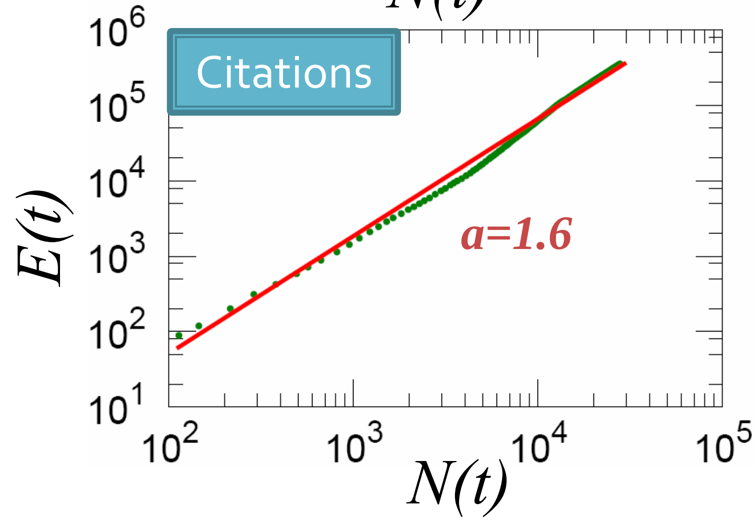
Q1) Network evolution

- What is the relation between the number of nodes and the edges over time?
- ~~Prior work assumes: constant average degree over time~~
- Networks are denser over time
- **Densification Power Law:**

$$E(t) \propto N(t)^a$$

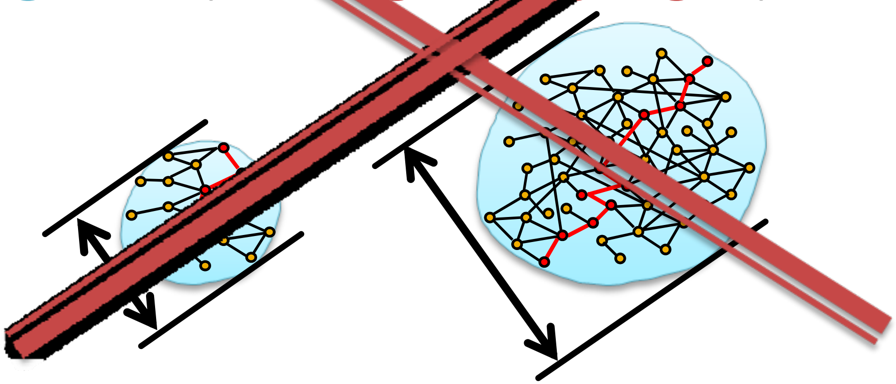
a … densification exponent (1 ≤ a ≤ 2)



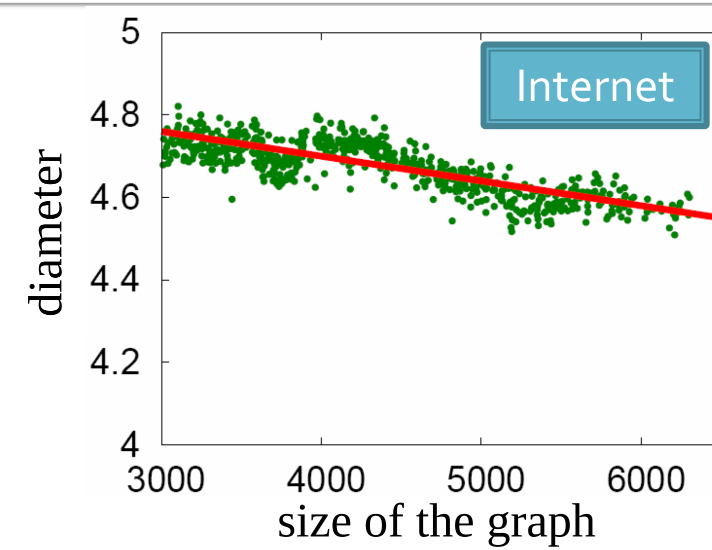


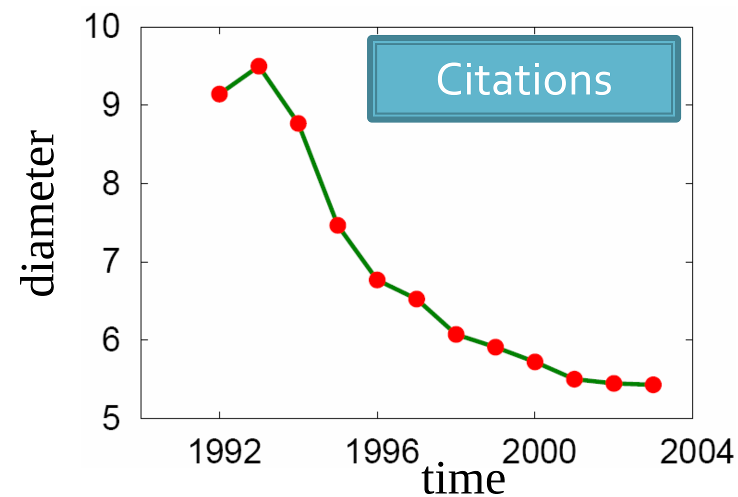
Q1) Network evolution

- ~~Prior models and intuition say that the network diameter slowly grows (like log N, log log N)~~



- **Diameter shrinks over time**
  - as the network grows the distances between the nodes slowly **decrease**

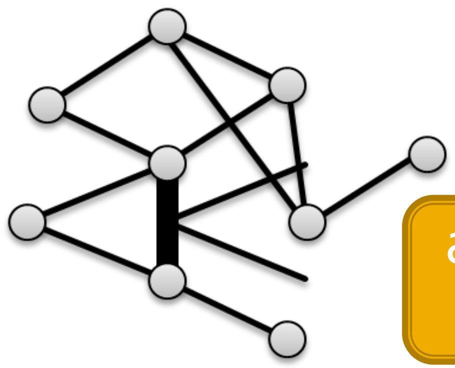




[w/ Backstrom-Kumar-Tomkins, KDD '08]

Q2) Modeling edge attachment

- We directly observe atomic events of network evolution (and not only network snapshots)



and so on for millions…

We can model evolution at finest scale

- Test individual edge attachment
  - Directly observe events leading to network properties
- Compare network models by likelihood (and not by just summary network statistics)

[w/ Backstrom-Kumar-Tomkins, KDD '08]

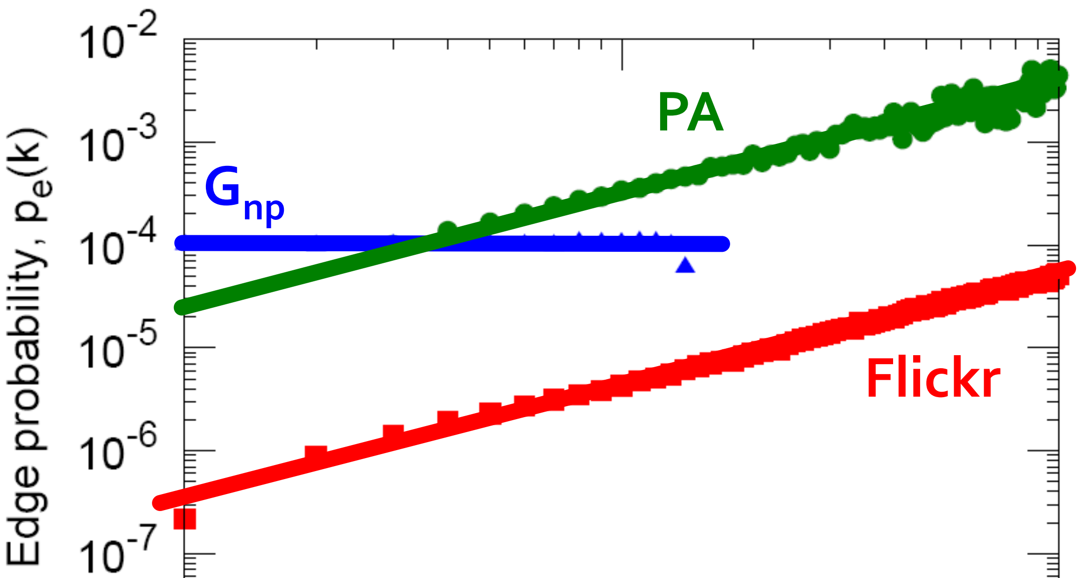
Setting: Edge-by-edge evolution

- Network datasets
  - Full temporal information from the first edge onwards
  - LinkedIn (N=7m, E=30m), Flickr (N=600k, E=3m), Delicious (N=200k, E=430k), Answers (N=600k, E=2m)
- We model 3 processes governing the evolution
  - P1) Node arrival: node enters the network
  - P2) Edge initiation: node wakes up, initiates an edge, goes to sleep
  - P3) Edge destination: where to attach a new edge
    - Are edges more likely to attach to high degree nodes?
    - Are edges more likely to attach to nodes that are close?

[w/ Backstrom-Kumar-Tomkins, KDD '08]

Edge attachment degree bias

- Are edges more likely to connect to higher degree nodes?



First direct proof of preferential attachment!

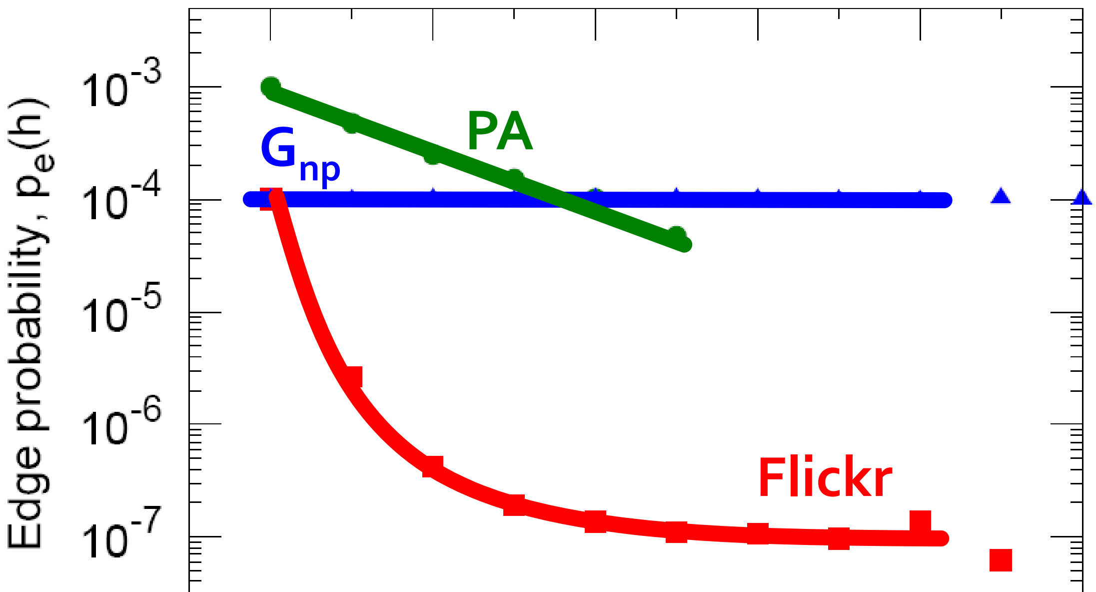
$$p_e(k) \propto k^{\tau}$$

Network	τ
G_{np}	0
PA	1
Flickr	1
Delicious	1
Answers	0.9
LinkedIn	0.6

[w/ Backstrom-Kumar-Tomkins, KDD '08]

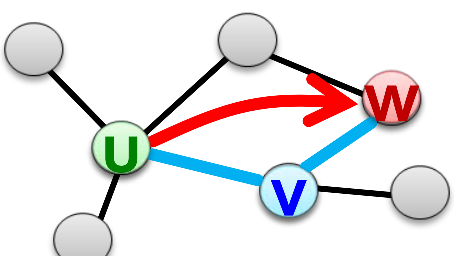
But, edges also attach locally

- Just before the edge *(u,w)* is placed how many hops is between *u* and *w*?



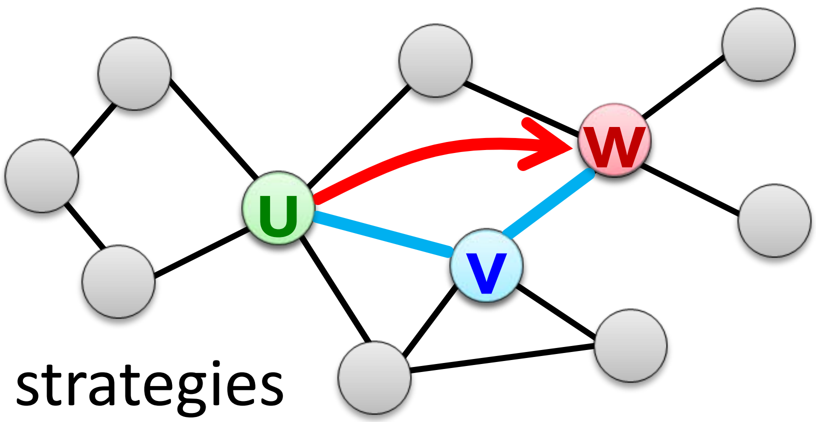
Real edges are local.
Most of them close triangles!

Fraction of triad closing edges	
Network	% Δ
Flickr	66%
Delicious	28%
Answers	23%
LinkedIn	50%



How to best close a triangle?

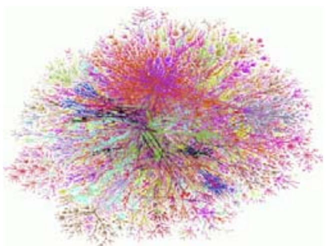
- New triad-closing edge *(u,w)* appears next
- We model this as:
  1. *u* chooses neighbor *v*
  2. v chooses neighbor *w*
  3. Connect *(u,w)*
  - We consider 25 triad closing strategies
  - and compute their log-likelihood



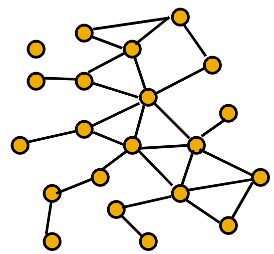
- Triad closing is best explained by
  - choosing a node based on the number of common friends and time since last activity
  - (just choosing random neighbor also works well)

Q3) Generating realistic graphs

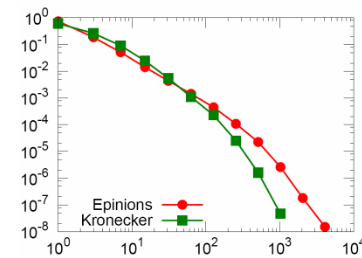
Problem: generate a realistic looking synthetic network



Given a real network



Generate a synthetic network

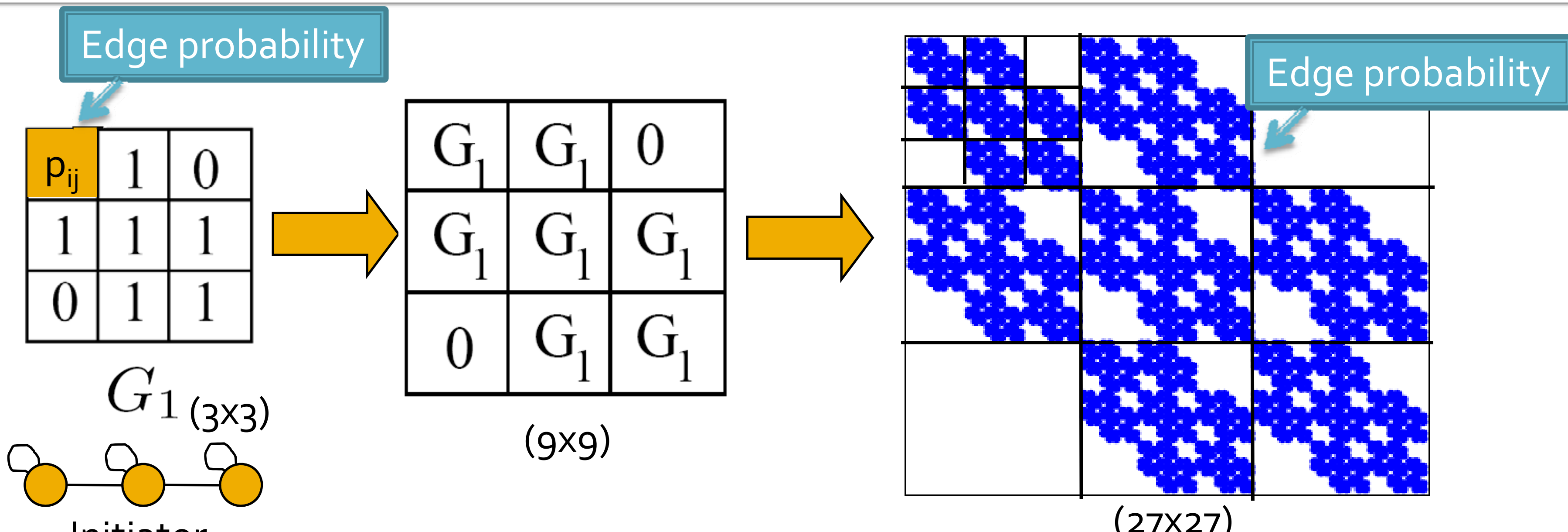


Compare graph properties, *e.g.*, degree distribution

- Why synthetic graphs?
  - Anomaly detection, Simulations, Predictions, Null-model, Sharing privacy sensitive graphs, …
- **Q:** Which network properties do we care about?
- **Q:** What is a good model and how do we **fit** it?

[w/ Chakrabarti-Kleinberg-Faloutsos, PKDD '05]

Q3) The model: Kronecker graphs



Kr... ...es

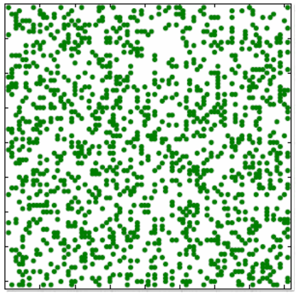
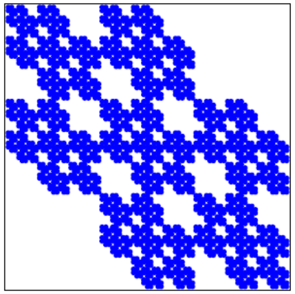
- W... ...hs:

Given a real graph.
How to estimate the initiator G_1?

Shrinking/stabilizing diameter, Spectral properties

Q5) Kronecker graphs: Estimation

- Maximum likelihood estimation

$\arg\max_{\Theta} P($  $|$  $\xleftarrow{\text{Kronecker}} \Theta)$

$$\Theta = \begin{array}{|c|c|} \hline a & b \\ \hline c & d \\ \hline \end{array}$$

- Naïve estimation takes *O(N!N²)*:
  - *N!* for diffe…
    - **Our solutio…** ***(big) const***
  - *N^2* for trave… …rix
    - **Our solutio…** …): *$N^2 \rightarrow E$*
- Do stochastic gradient descent

We estimate the model in **O(E)**

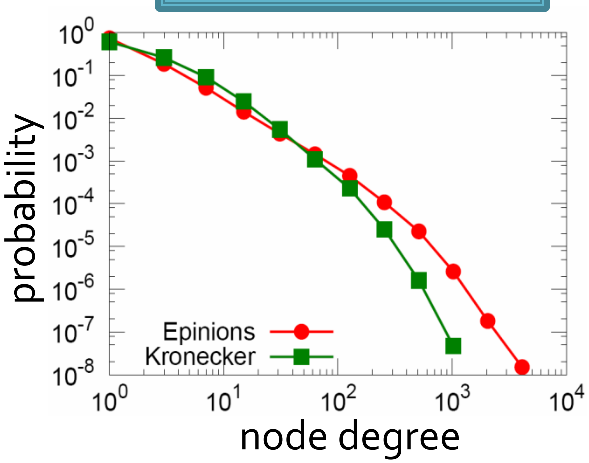
[w/ Faloutsos, ICML '07]

Estimation: Epinions (N=76k, E=510k)

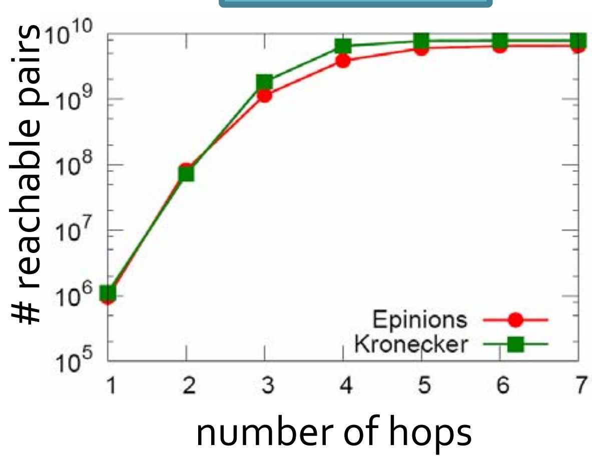
- We search the space of $\sim 10^{1,000,000}$ permutations
- Fitting takes 2 hours
- Real and Kronecker are very close

$$\hat{\Theta} = \begin{array}{|c|c|} \hline 0.99 & 0.54 \\ \hline 0.49 & 0.13 \\ \hline \end{array}$$

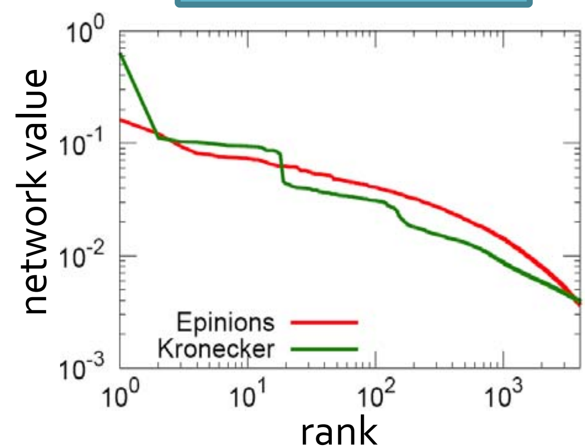
Degree distribution



Path lengths



"Network" values



Thesis: The structure

	Network Evolution	Network Cascades	Large Data
Observations	**Q1:** How does network structure evolve over time?	**Q4:** What are patterns of diffusion in networks?	**Q7:** What are the properties of a social network of the whole planet?
Models	**Q2:** How to model individual edge attachment?	**Q5:** How do we model influence propagation?	**Q8:** What is community structure of large networks?
Algorithms (applications)	**Q3:** How to generate realistic looking networks?	**Q6:** How to identify influential nodes and epidemics?	**Q9:** How to predict search result quality from the web graph?

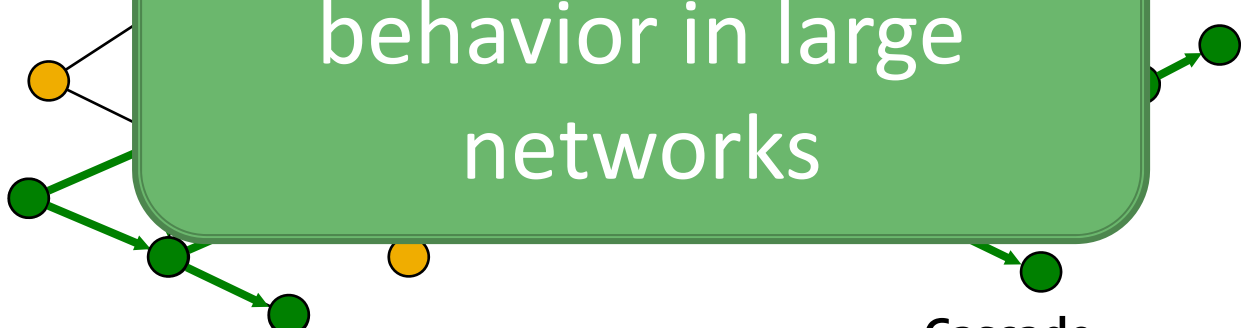
Thesis: The structure

	Network Evolution	Network Cascades	Large Data
Observations	**Q1:** How does network structure evolve over time?	**Q4:** What are patterns of diffusion in networks?	**Q7:** What are the properties of a social network of the whole planet?
Models	**Q2:** How to model individual edge attachment?	**Q5:** How can we model influence propagation?	**Q8:** What is community structure of large networks?
Algorithms (applications)	**Q3:** How to generate realistic looking networks?	**Q6:** How to identify influential nodes and epidemics?	**Q9:** How to predict search result quality from the web graph?

Part 2: Diffusion and Cascades

- Behavior that cascades from node to node like an epidemic
  - News, opinions, rumors
  - Word-of-mouth in marketing
  - Infectious diseases
- As activati… …ork they leav…

We observe cascading behavior in large networks



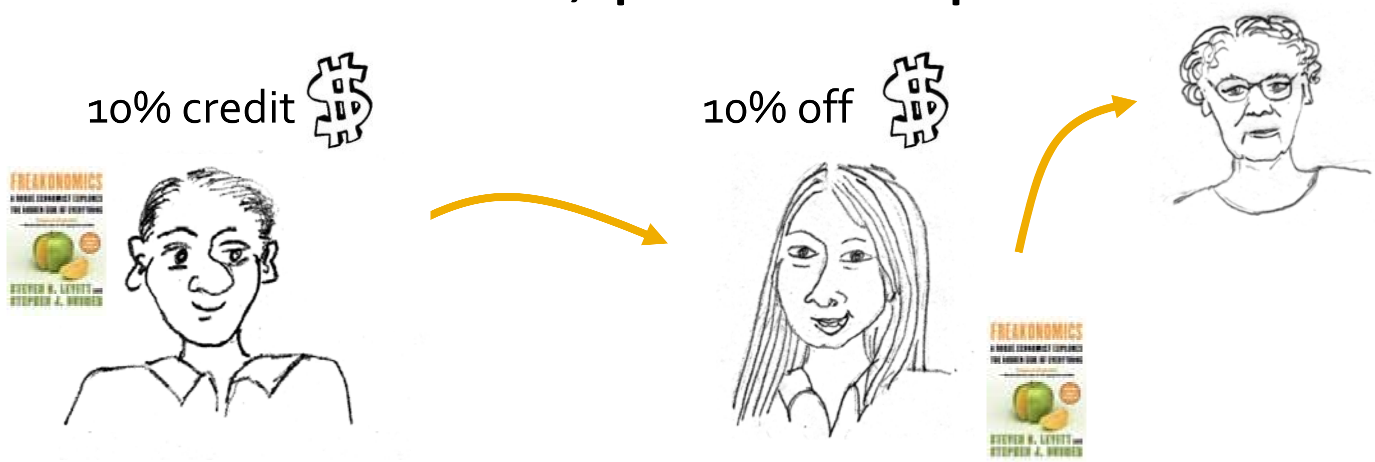
Network

Cascade
(propagation graph)

[w/ Adamic-Huberman, EC ’06]

Setting 1: Viral marketing

- People send and receive product recommendations, purchase products

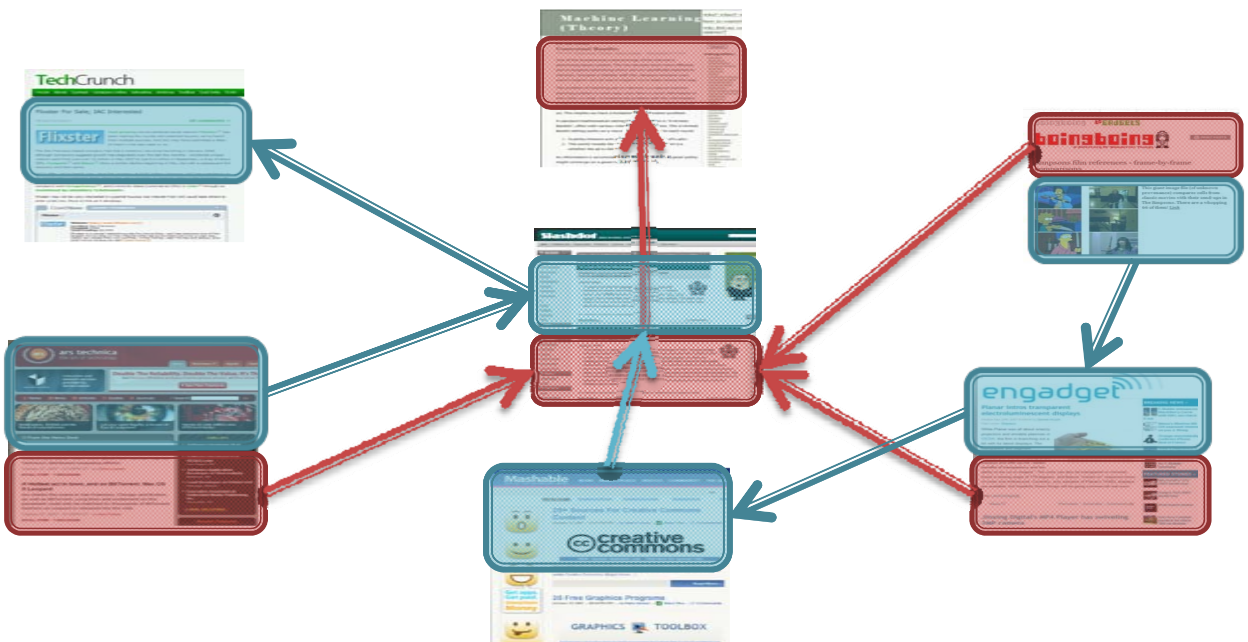


- Data: Large online retailer: 4 million people, 16 million recommendations, 500k products

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

Setting 2: Blogosphere

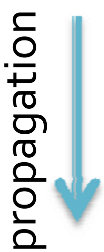
- Bloggers write posts and refer (link) to other posts and the information propagates



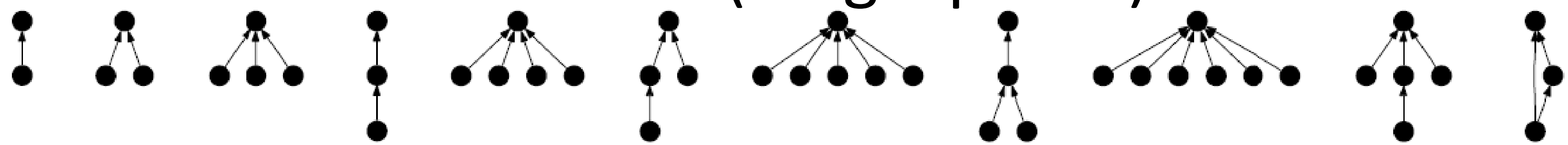
- Data: 10.5 million posts, 16 million links

[w/ Kleinberg-Singh, PAKDD '06]

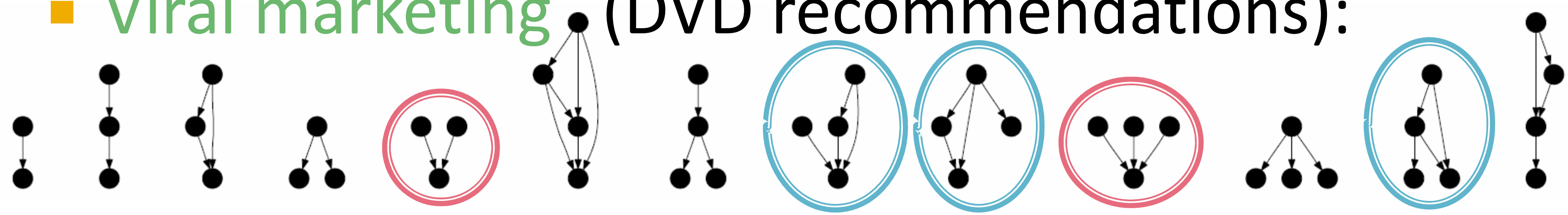
Q4) What do cascades look like?



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



- Viral marketing (DVD recommendations):

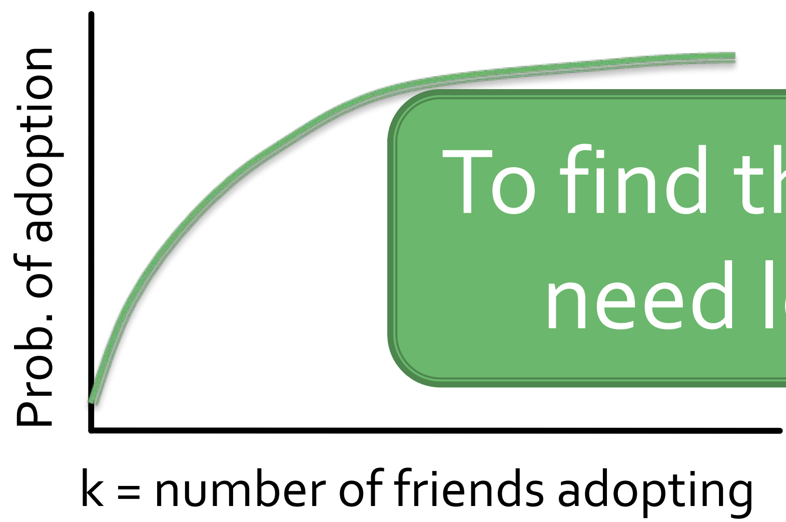


(ordered by frequency)

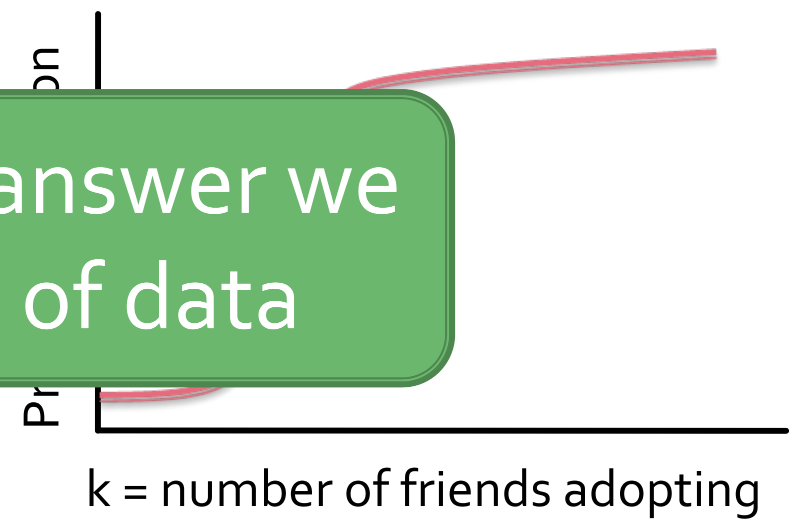
- Viral marketing cascades are more social:
  - Collisions (no summarizers)
  - Richer non-tree structures

Q5) 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



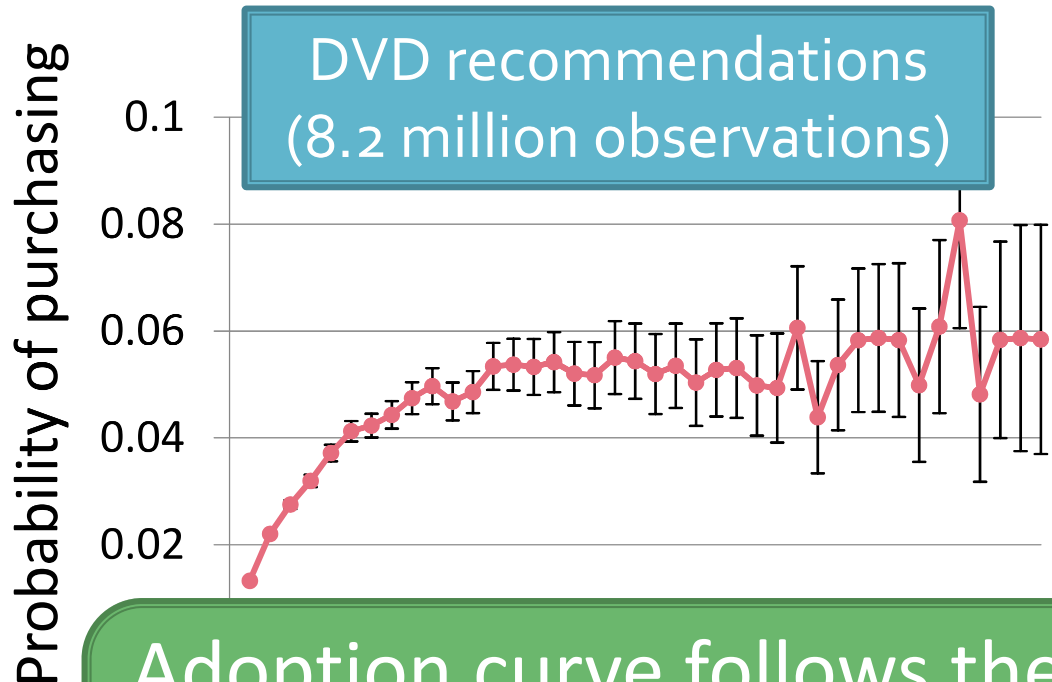
Diminishing returns?



Critical mass?

To find the answer we need lots of data

Q5) Adoption curve: Validation

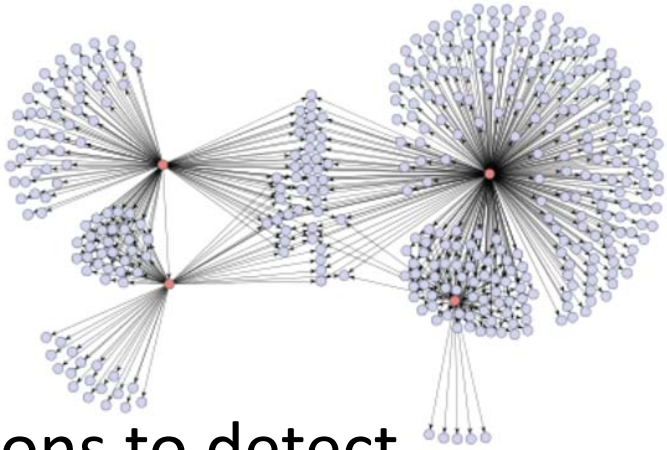


Adoption curve follows the **diminishing returns.**
Can we exploit this?

Later similar fin… -Huttenlocher-
Kleinberg ‘06], and probability of communication [Kossinets-Watts ’06]

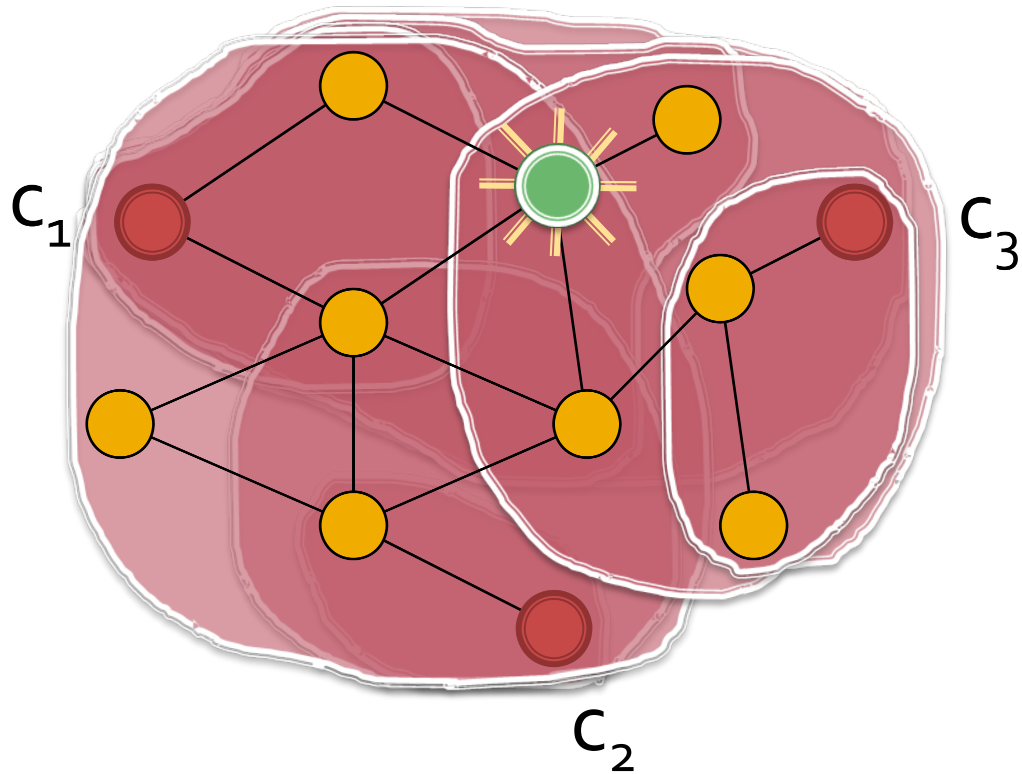
Q6) Cascade & outbreak detection

- 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]

Q6) The problem: Detecting cascades

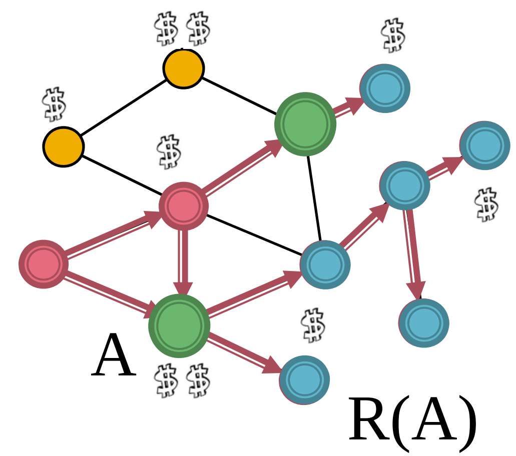


How to quickly detect epidemics as they spread?

[w/ Krause-Guestrin et al., KDD '07]

Two parts to the problem

- Cost:
  - Cost of monitoring is node dependent
- Reward:
  - Minimize the number of affected nodes:
    - If A are the monitored nodes, let R(A) denote the number of nodes we save



(We also consider other rewards:
- Minimize time to detection
- Maximize number of detected outbreaks)

Optimization problem

- Given:
  - Graph G(V,E), budget M
  - Data on how cascades $C_1, \ldots, C_i, \ldots, C_K$ spread over time
- Select a set of nodes A maximizing the reward

$$\max_{A \subseteq V} \sum_i \mathrm{Prob}(i) \underbrace{R_i(A)}_{\text{Reward for detecting cascade } i}$$

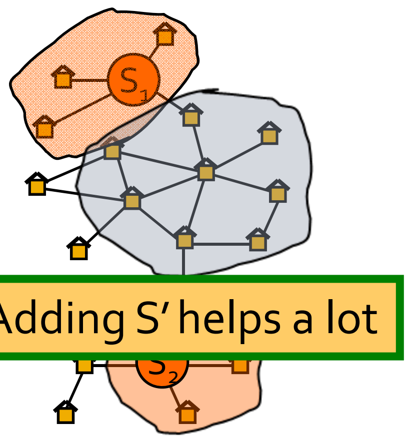
subject to $cost(A) \leq M$

- Solving the problem exactly is NP-hard
  - Max-cover [Khuller et al. '99]

Solution: CELF Algorithm

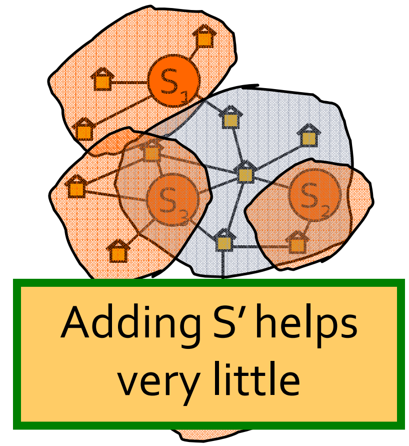
- We develop CELF (cost-effective lazy forward-selection) algorithm:
  - Two independent runs of a modified greedy
    - Solution set A': ignore cost, greedily optimize reward
    - Solution set A'': greedily optimize reward/cost ratio
  - Pick best of the two: *arg max(R(A'), R(A''))*
- Theorem: If R is submodular then CELF is near optimal
  - CELF achieves *½(1-1/e)* factor approximation

Problem structure: Submodularity

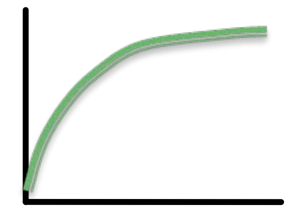


Placement A={S_1, S_2}

New monitored node:



Placement B={S_1, S_2, S_3, S_4}

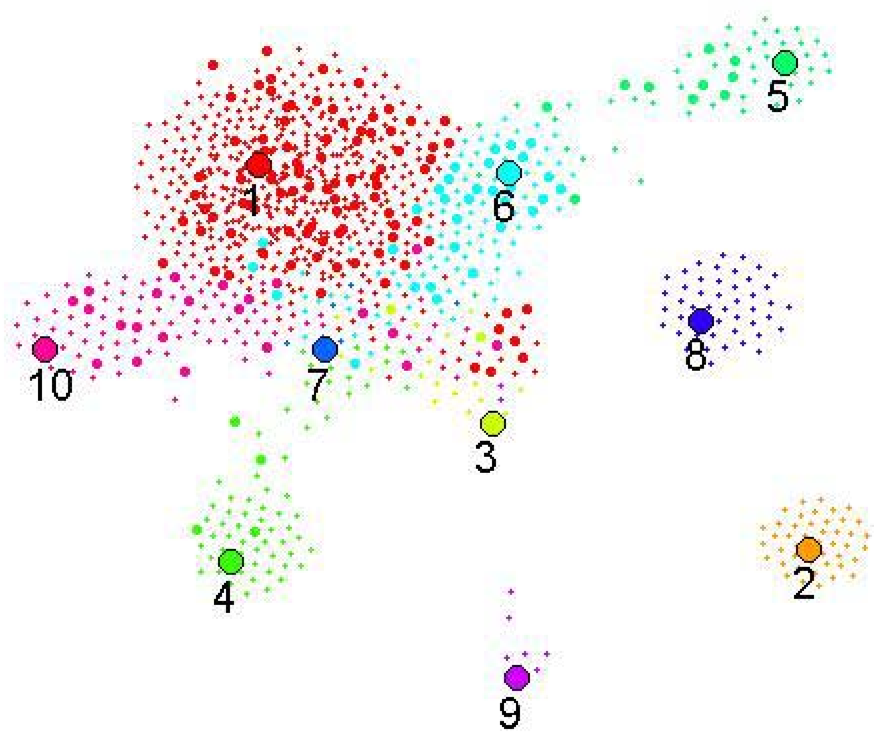


- Theorem: Reward function *R* is submodular (diminishing returns, think of it as "concavity")

$$\underbrace{R(A \cup \{u\}) - R(A)}_{\text{Gain of adding a node to a small set}} \geq \underbrace{R(B \cup \{u\}) - R(B)}_{\text{Gain of adding a node to a large set}} \qquad A \subseteq B$$

Blogs: Information epidemics

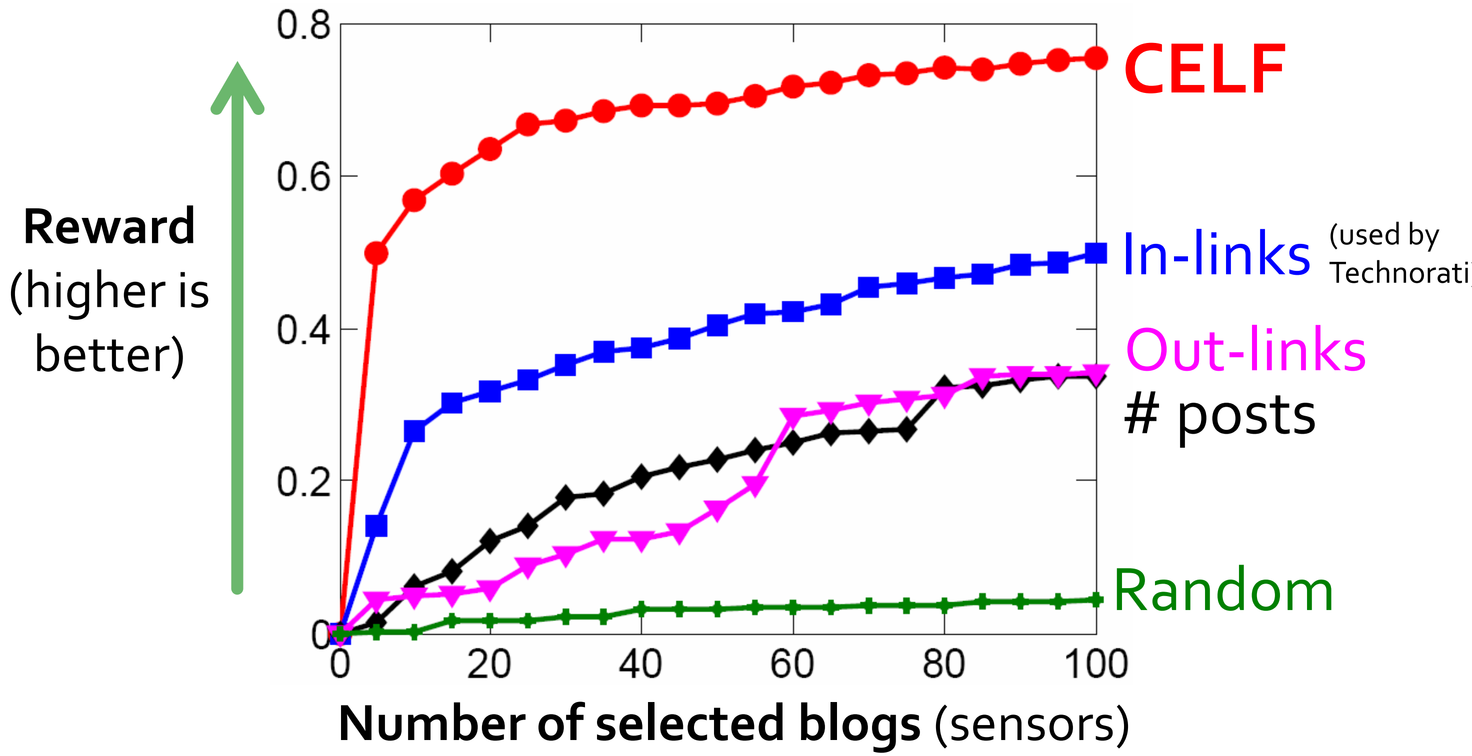
- **Question:** Which blogs should one read to catch big stories?
- **Idea:** Each blog covers part of the blogosphere



- Each dot is a blog
- Proximity is based on the number of common cascades

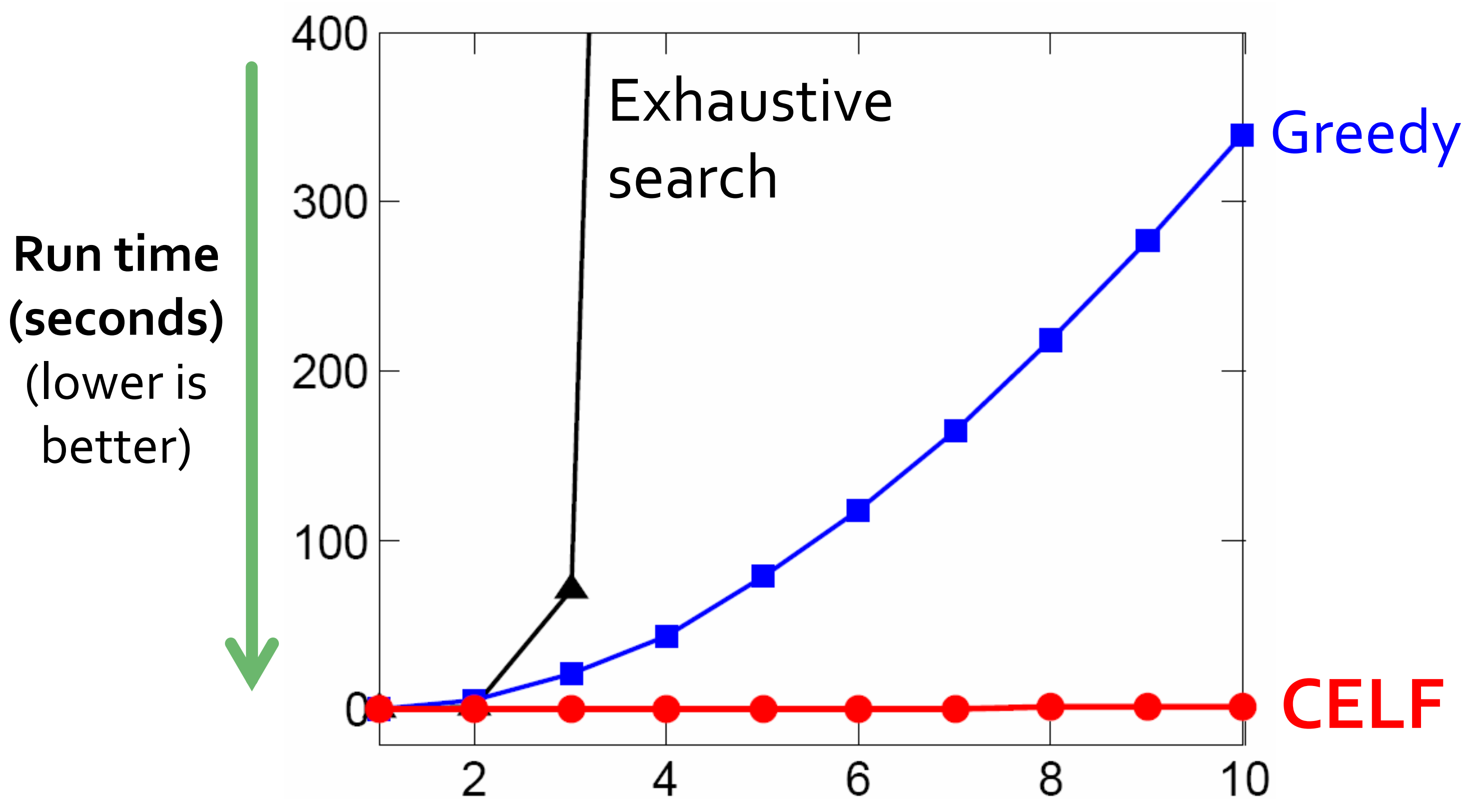
Blogs: Information epidemics

- Which blogs should one read to catch big stories?



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

CELF: Scalability

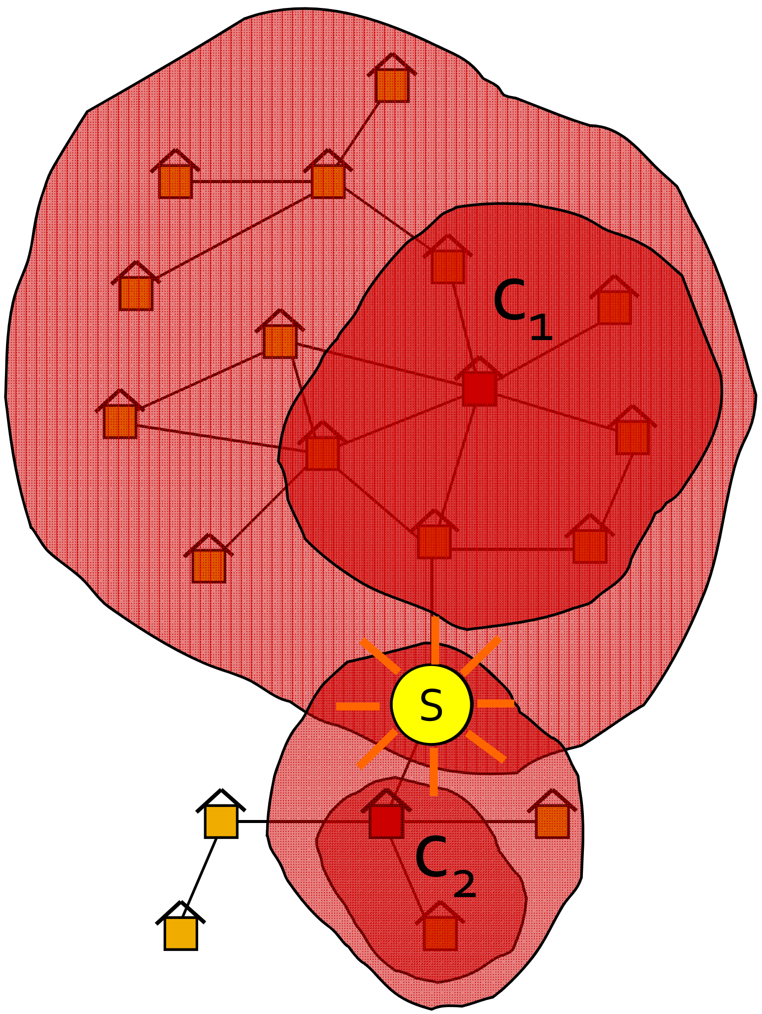


CELF runs **700x** faster than simple greedy algorithm

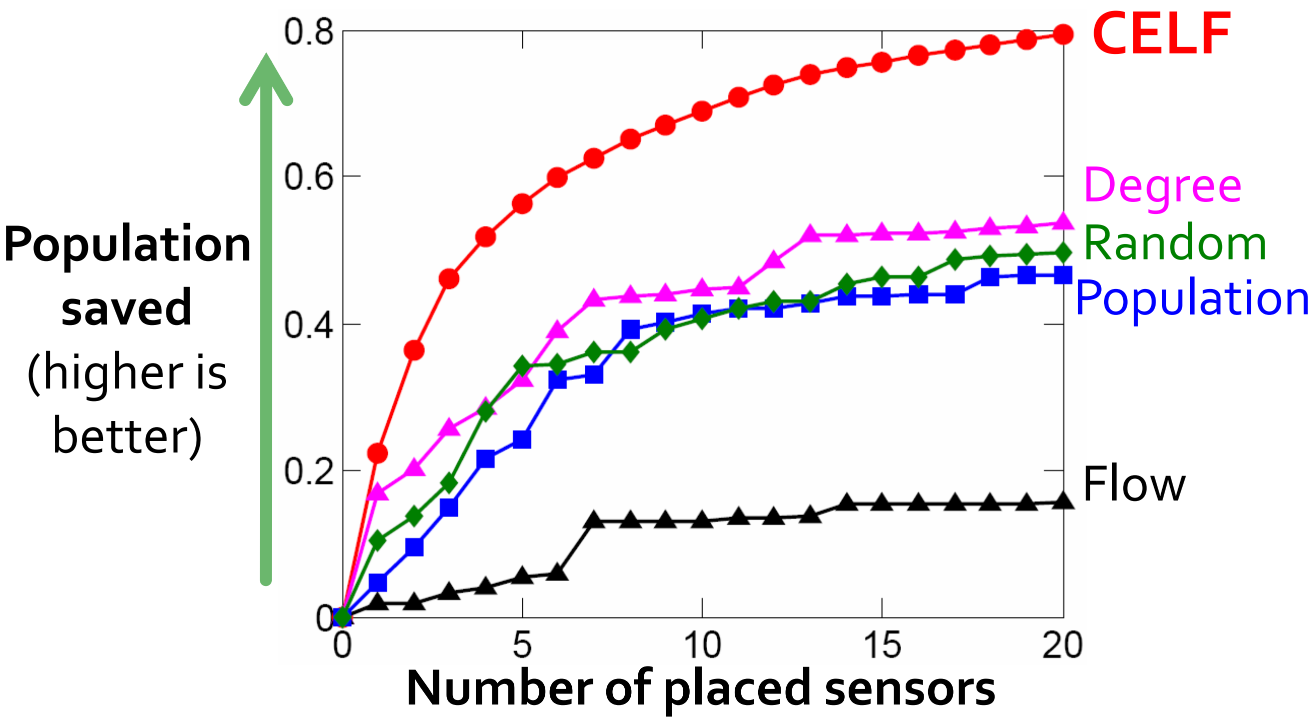
[w/ Krause et al., J. of Water Resource Planning]

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*



Water network: Results



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

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

Thesis: The structure

	Network Evolution	Network Cascades	Large Data
Observations	**Q1:** How does network structure evolve over time?	**Q4:** What are patterns of diffusion in networks?	**Q7:** What are the properties of a social network of the whole planet?
Models	**Q2:** How to model individual edge attachment?	**Q5:** How do we model influence propagation?	**Q8:** What is community structure of large networks?
Algorithms (applications)	**Q3:** How to generate realistic looking networks?	**Q6:** How to identify influential nodes and epidemics?	**Q9:** How to predict search result quality from the web graph?

Thesis: The structure

	Network Evolution	Network Cascades	Large Data
Observations	**Q1:** How does network structure evolve over time?	**Q4:** What are patterns of diffusion in networks?	**Q7:** What are the properties of a social network of the whole planet?
Models	**Q2:** How to model individual edge attachment?	**Q5:** How do we model influence propagation?	**Q8:** What is community structure of large networks?
Algorithms (applications)	**Q3:** How to generate realistic looking networks?	**Q6:** How to identify influential nodes and epidemics?	**Q9:** How to predict search result quality from the web graph?

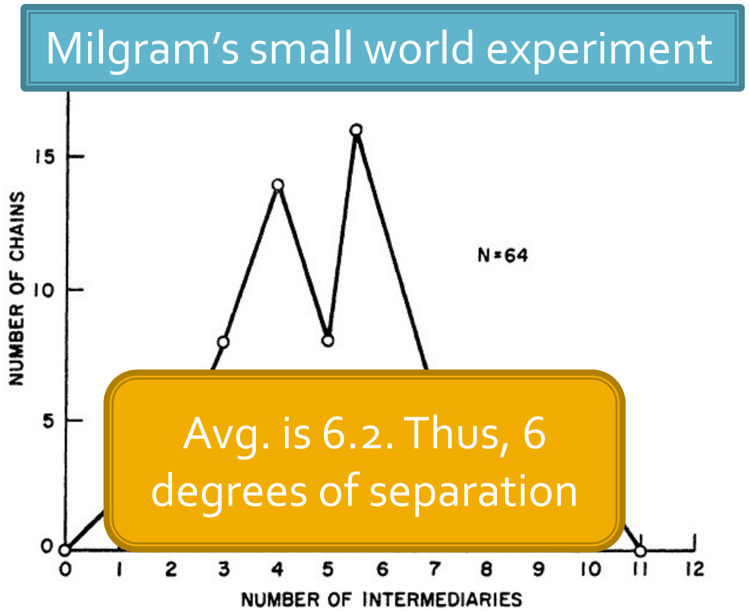
3 case studies on large data

Benefits from working with large data:

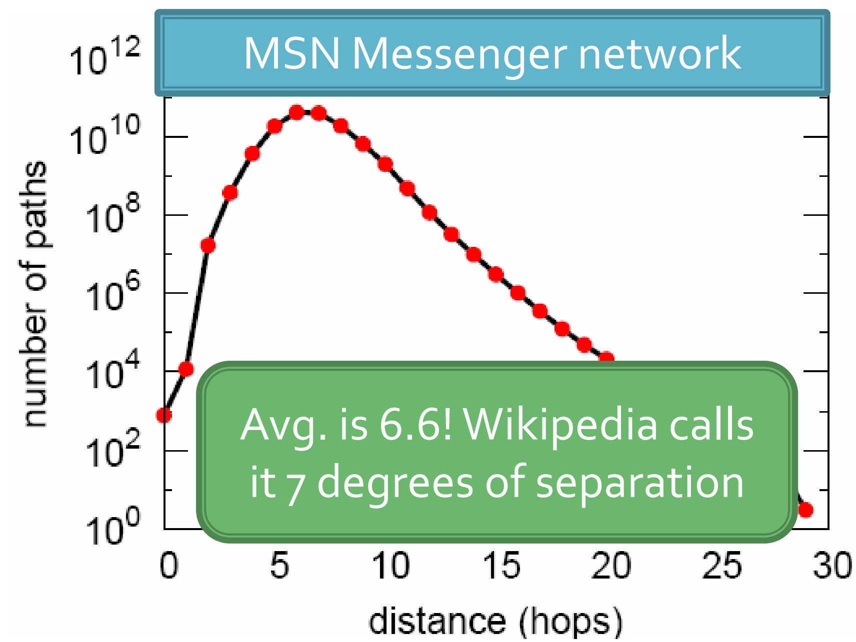
- **Q7)** Can test hypothesis at planetary scale
  - 6 degrees of separation
- **Q8)** Observe phenomena previously invisible
  - Network community structure
- **Q9)** Making global predictions from local network structure
  - Web search

Q7) Planetary look on a small-world

- Small-world experiment [Milgram ‘67]
  - People send letters from Nebraska to Boston
- How many steps does it take?
- **Messenger social network** – largest network analyzed
  - 240M people, 255B messages, 4.5TB data

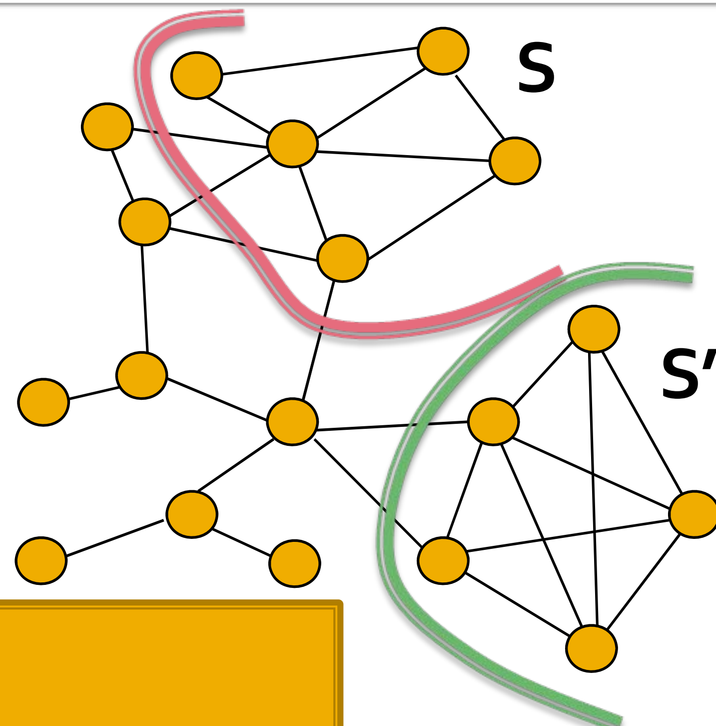


(i.e., hops + 1)



Q8) Network community structure

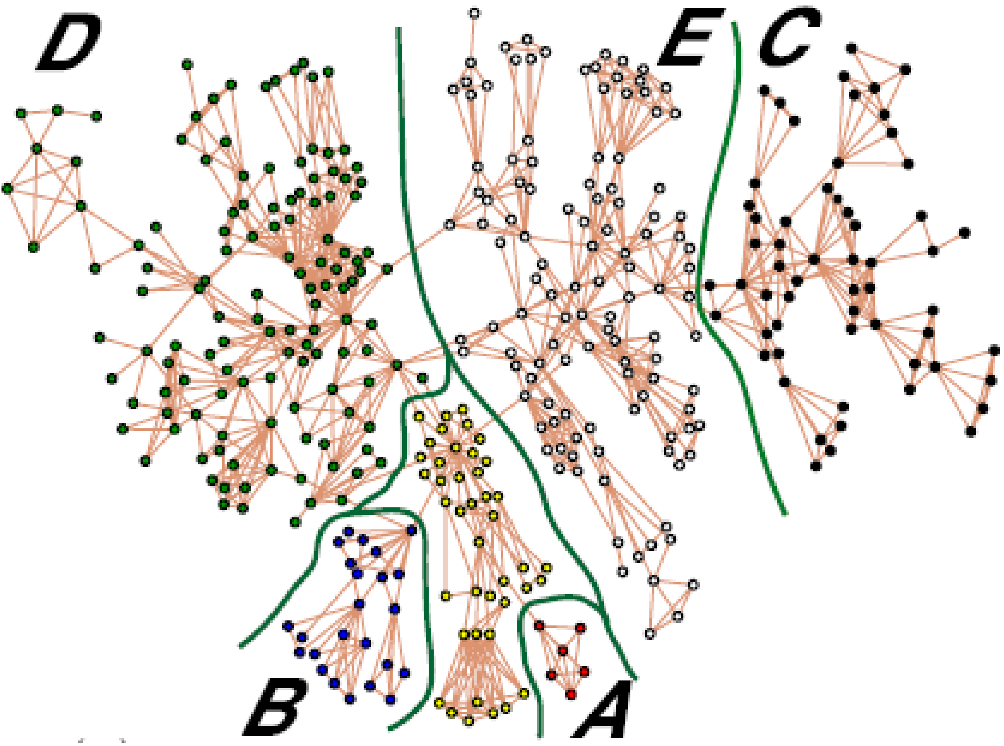
- How community like is a set of nodes?
- Need a natural intuitive measure

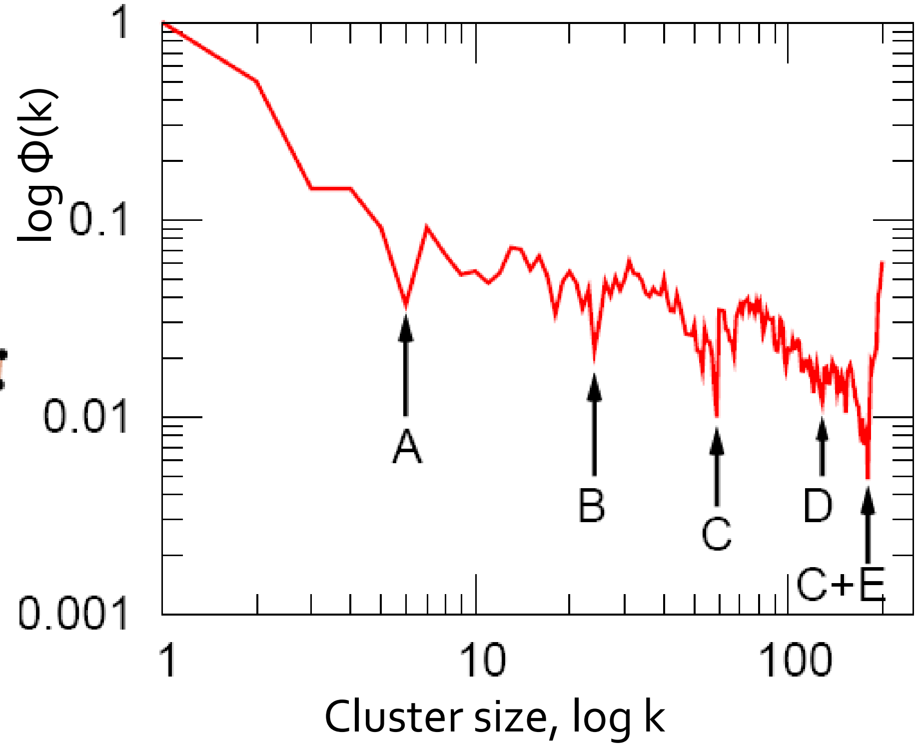


Conductance:
$\Phi(S)$ = # edges cut / # edges inside

- **Plot:** Score of best cut of volume $k=|S|$

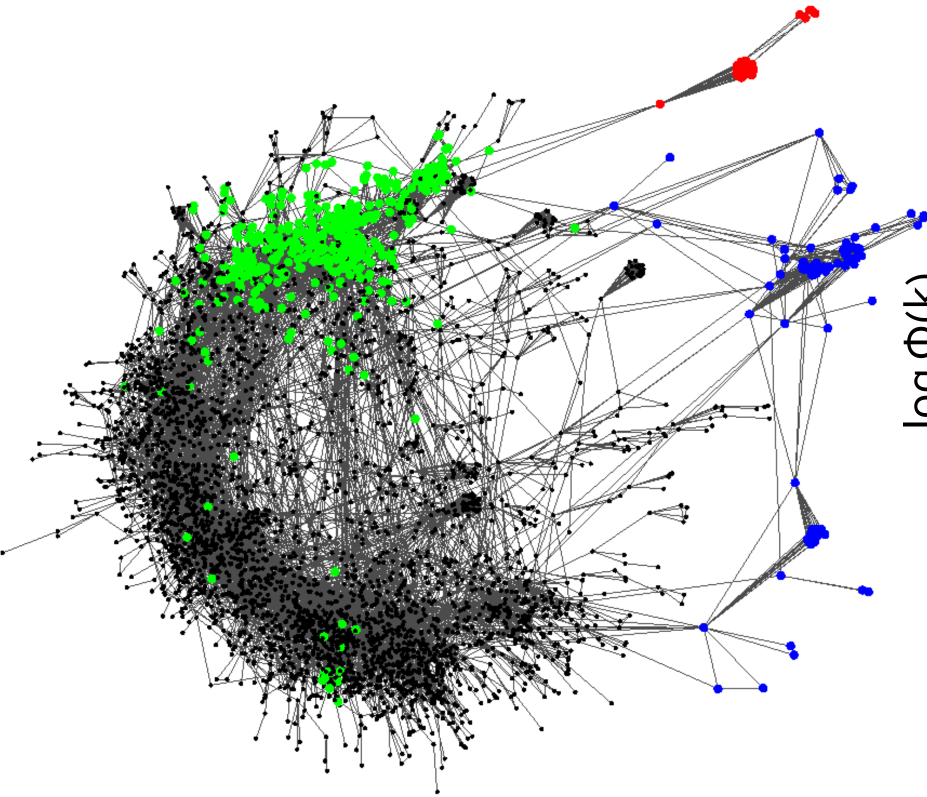
Example: Small network

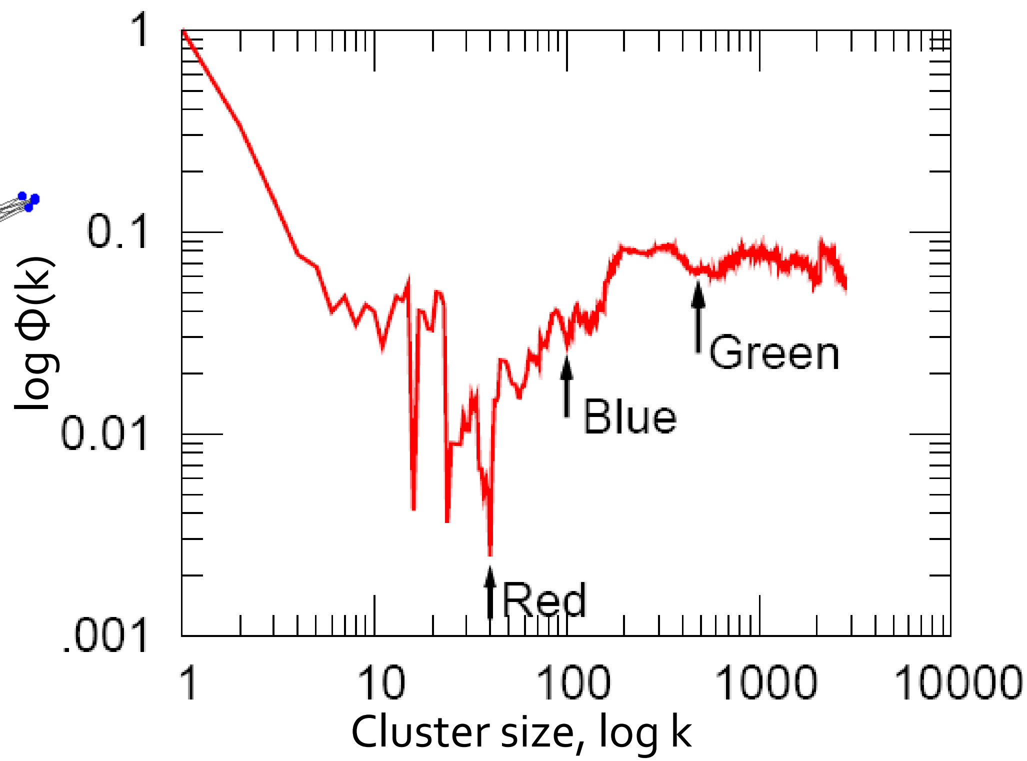




Collaborations between scientists (N=397, E=914)

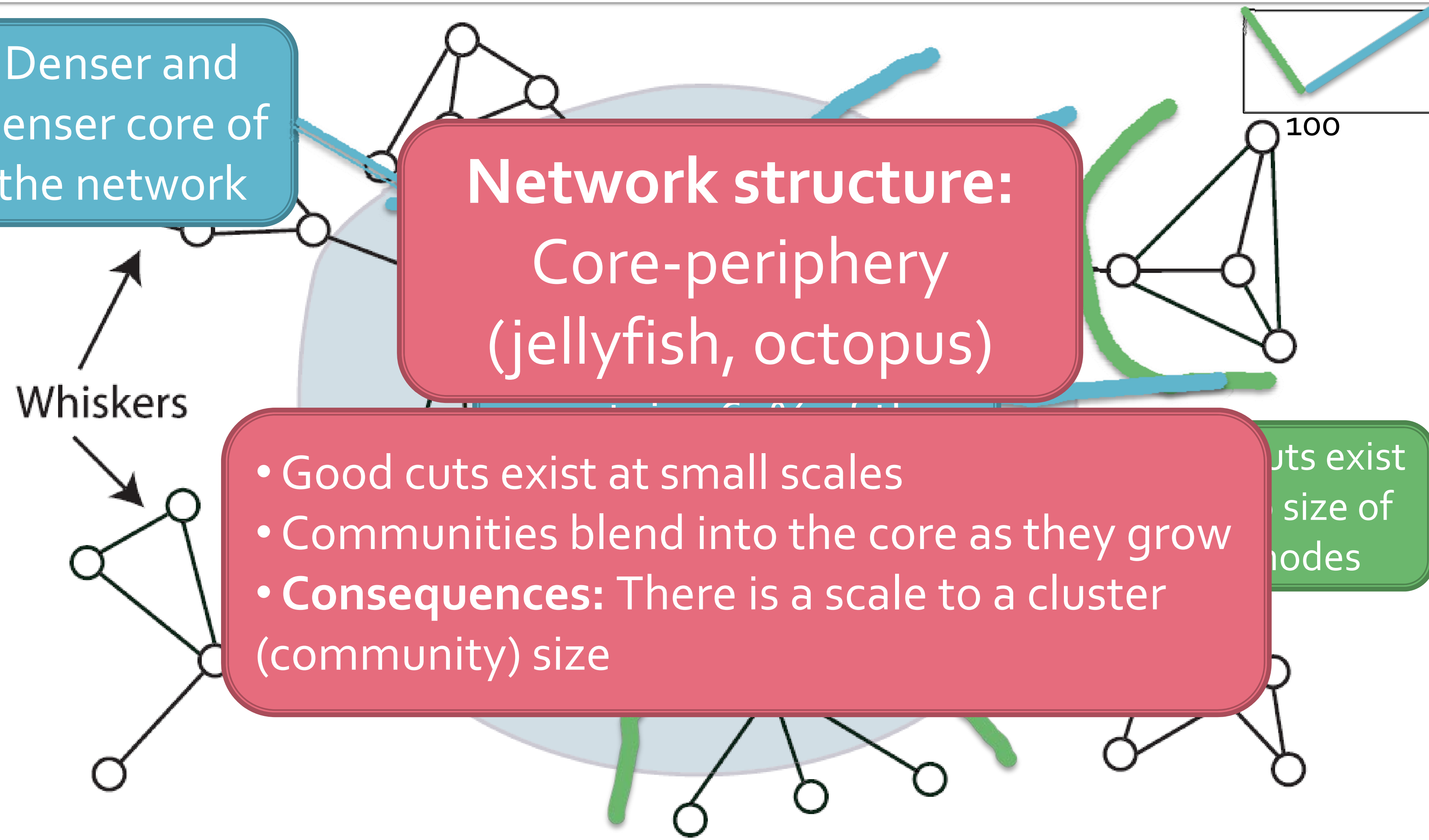
Example: Large network





Collaboration network (N=4,158, E=13,422)

Q8) Suggested network structure



Denser and denser core of the network

Network structure:
Core-periphery
(jellyfish, octopus)

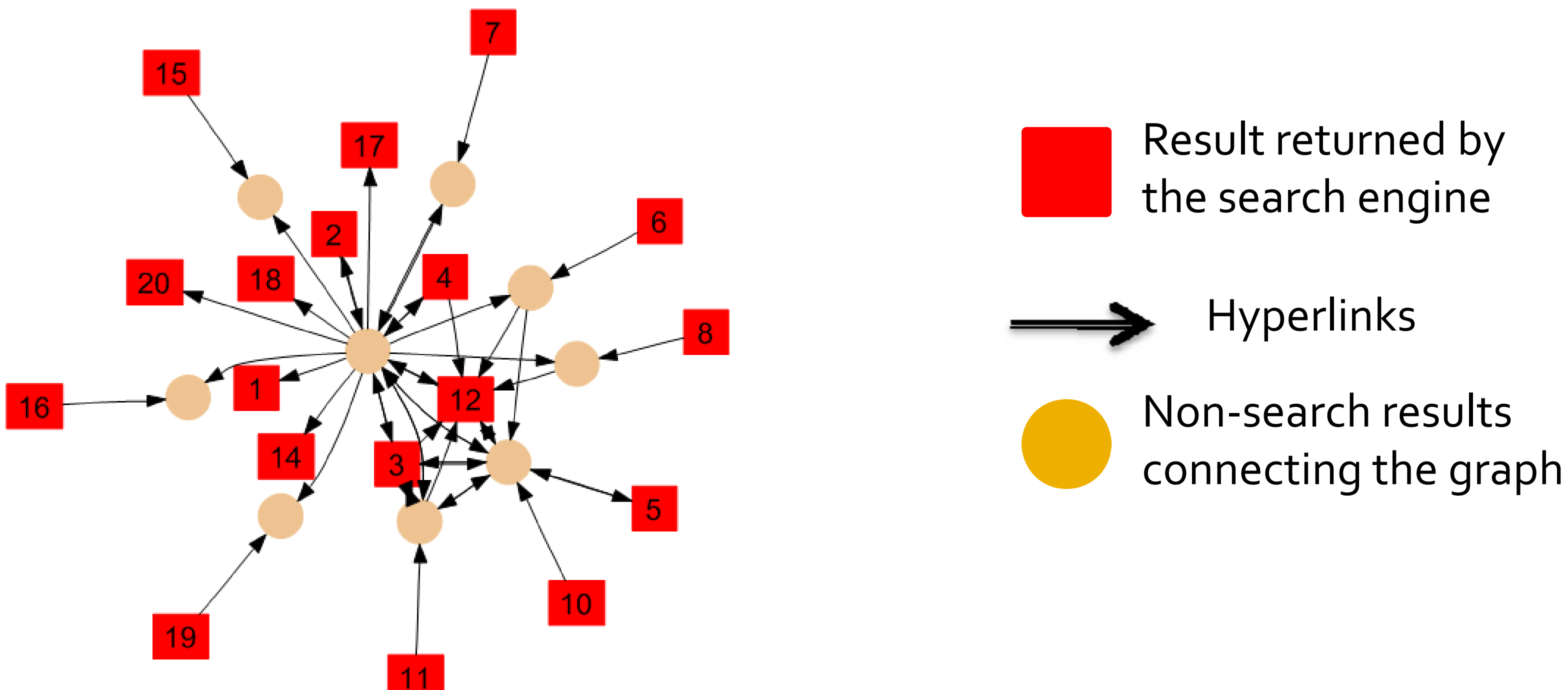
- Good cuts exist at small scales
- Communities blend into the core as they grow
- **Consequences:** There is a scale to a cluster (community) size

uts exist
size of
odes

[w/ Dumais-Horvitz, WWW '07]

Q9) Web Projections

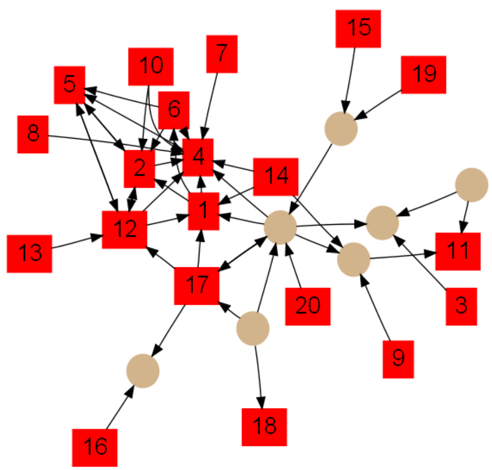
- User types in a query to a search engine
- Search engine returns results:



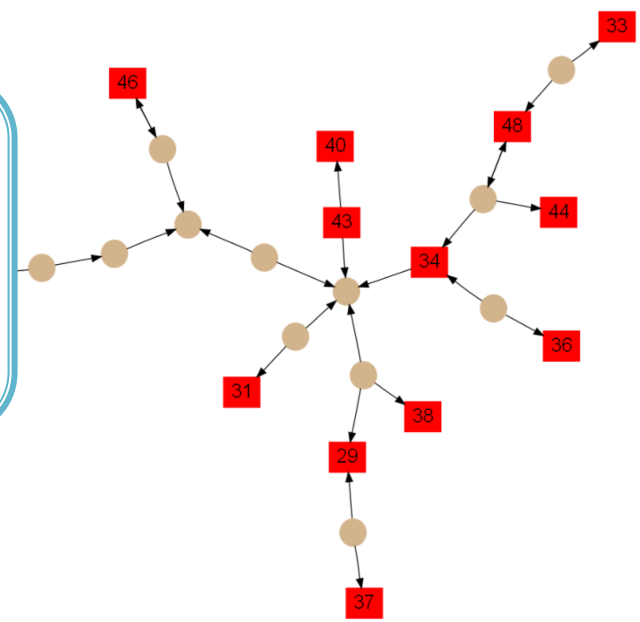
Is this a good set of search results?

Q9) Web Projections: Results

- We can predict search result quality with **80% accuracy** just from the connection patterns between the results



Good?
Poor?



Predict "**Good**"

Predict "**Poor**"

Thesis: The structure

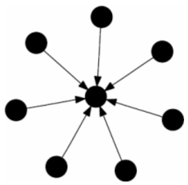
	Network Evolution	Network Cascades	Large Data
Observations	Densification and shrinking diameter	Cascade shapes	7 degrees of separation of MSN
Models	Triangle closing model	Diminishing returns of human adoption	Network community structure
Algorithms (applications)	Kronecker graphs and fitting	Cascade and outbreak detection	Web projections

Future directions: Evolution

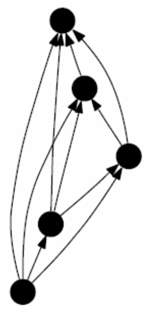
- Why are networks the way they are?
- Health of a social network
  - Steer the network evolution
  - Better design networked services
- Predictive modeling of large communities
  - Online massively multi-player games are closed worlds with detailed traces of activity

Future directions: Diffusion

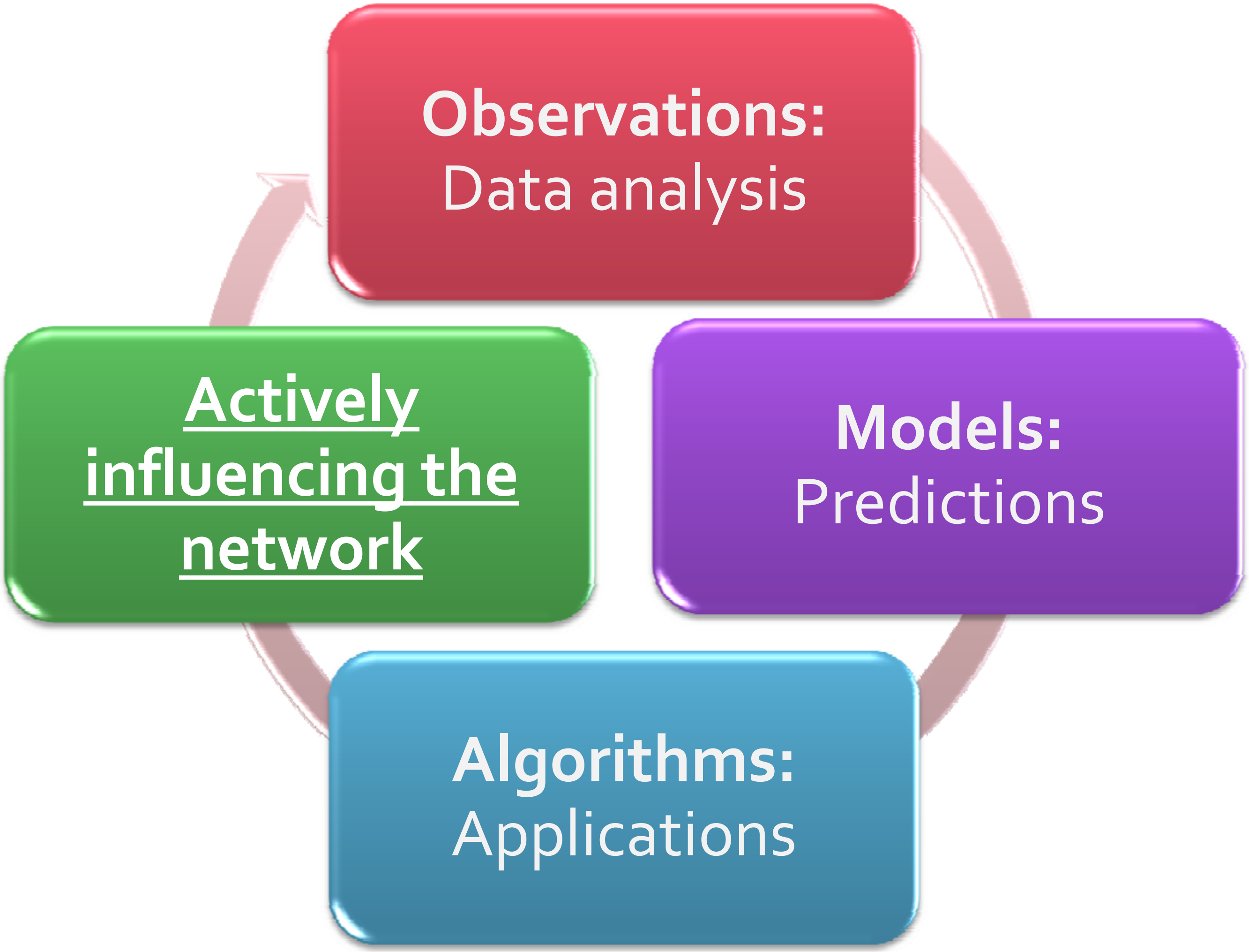
- Predictive models of information diffusion
  - When, where and what post will create a cascade?
  - Where should one tap the network to get the effect they want?
  - Social Media Marketing



- How do news and information spread
  - New ranking and influence measures for blogs
  - Sentiment analysis from cascade structure



What's next?


Observations:
Data analysis
Models:
Predictions
Algorithms:
Applications
Actively influencing the network

Thanks!

Everyone is invited to
Wean 4623 at 5pm
There will be food ☺

Thesis: The structure

	Network Evolution	Network Cascades	Large Data
Observations	Densification and shrinking diameter	Cascade shapes	7 degrees of separation of MSN
Models	Triangle closing model	Diminishing returns of human adoption	Network community structure
Algorithms (applications)	Kronecker graphs and fitting	Cascade and outbreak detection	Web projections