Social Media Analytics:
Part 1: Information flow
Information and Networks

- Information reaches us...
  - ...by personal influence in our social networks
  - ...through transmission by mass media
- Social Media is media designed to be disseminated through Social interaction
  - How does information transmitted by the media interact with the personal influence arising from social networks?
  - Tension between global effects from the mass media and local effects carried by social structure
Social Media: Big change

- Web is no longer a static library that people passively browse
- Web is a place where people:
  - Consume and create content
  - Interact with other people:
    - Internet forums, Blogs, Social networks, Twitter, Wikis, Podcasts, Slide sharing, Bookmark sharing, Product reviews, Comments, ...
- Facebook traffic tops Google (for USA)
  - March 2010: FB > 7% of US traffic

http://money.cnn.com/2010/03/16/technology/facebook_most_visited
Rich and big data:

- Billions users, billions contents
- Textual, Multimedia (image, videos, etc.)
- Billions of connections
- Behaviors, preferences, trends...

Data is open and easy to access

- It’s easy to get data from Social Media
- Datasets
- Developers APIs
- Spidering the Web
Social Media Datasets

- **Social Tagging:**
  - CiteULike, Bibsonomy, MovieLens, Delicious, Flickr, Last.FM...
    - [http://kmi.tugraz.at/staff/markus/datasets/](http://kmi.tugraz.at/staff/markus/datasets/)

- **Yahoo! Firehose**
  - 750K ratings/day, 8K reviews/day, 150K comments/day, status updates, Flickr, Delicious...

- **MySpace data** (real-time data, multimedia content, ...)

- **Spinn3r Blog Dataset, JDPA Sentiment Corpus**
  - [http://www.icwsm.org/data/](http://www.icwsm.org/data/)

For the list of datasets see tutorial website:
and also: [http://snap.stanford.edu/data](http://snap.stanford.edu/data)
Social Media: Opportunities

- Any user can share and contribute content, express opinions, link to others
- This means: Can data-mine opinions and behaviors of millions of users to gain insights into:
  - Human behavior
  - Marketing analytics
  - Product sentiment
Social Media: Value proposition

Consumer Generated, Not Edited, Not Authenticated

Actionable Intelligence
Applications: Reputation management

- Consumer Brand Analytics
  - What are people saying about our brand?
- Marketing Communications
  - Significant spending on marketing, advertising: Companies trying to position their products
  - Brand analytics helps to determine whether such campaigns are effective
- Product reviews
  - Automatically mine product reviews for information on product features, new requests, ...
    - Easy to use, Comfortable chair, Light weight, Sturdy, Good price
Applications: Citizen response

- **Citizen response**
  - solicit citizen feedback on bills debated in Congress
  - What new issues are being raised, what aspects of bill are popular, unpopular

- **Political Campaigns**
  - Why do people support a candidate?

- **Law enforcement**
  - Gang members boast about their activities on Facebook
  - Protests being planned through Twitter
  - NYT: Sending the Police Before There’s a Crime
    

8/21/2011
Citizen journalism provides more valuable information than newswire services

Challenge:
- Many redundant posts, users have to wade through hundreds of posts to locate useful information

Goal:
- Mine this data in real-time and produce well organized summaries
Viral marketing:
  - Personalized recommendations

Online forum users are brand advocates:
  - 79.2% of forum contributors help a friend to make a decision about a product purchase (47.6% of non-contributors).
  - 65% of forum contributors share advice (offline and in person) based on information that they’ve read online (35% of non-contributors)

http://www.socialmediaexaminer.com/new-studies-show-value-of-social-media
Process social media content, provide tools for analysts to:

- Identify social networks: groups, members
- Identify topics and sentiment
The tutorial: Social Media Analytics

- **Goal:** Introduce methods and algorithms for Social Media Analytics
- **Tutorial has two parts:**
  - **Part 1: Information Flow**
    - How do we capture and model the flow of information through networks to:
      - Predict information attention/popularity
      - Detect information big stories before they happen
  - **Part 2: Rich Interactions**
    - How do we go beyond “link”/“no-link”:
      - Predicting future links and their strengths
      - Separating friends from foes
Part 1: Information flow in networks

- 1.1: Data collection: How to track the flow?
- 1.2: Modeling and predicting the flow
- 1.3: Infer networks of information flow

Part 2: Rich interactions
Part 1 of the Tutorial: Overview

- Information flow through Social Media
  - Analyzing underlying mechanisms for the real-time spread of information through on-line networks

- Motivating questions:
  - How do messages spread through social networks?
  - How to predict the spread of information?
  - How to identify networks over which the messages spread?
Spinn3r Dataset: http://spinn3r.com

- 30 million articles/day (50GB of data)
  - 20,000 news sources + millions blogs and forums
  - And some Tweets and public Facebook posts

What are basic “units” of information?

- Pieces of information that propagate between the nodes (users, media sites, ...)
  - phrases, quotes, messages, links, tags
Would like to track units of information that:
- correspond to pieces of information:
  - events, articles, ...
- vary over the order of days,
- and can be handled at large scale

Ideas:
- (1) Cascading links to articles
- Textual fragments that travel relatively unchanged:
  - (2) URLs and hashtags on Twitter
  - (3) Phrases inside quotes: “...”
Bloggers write posts and refer (link) to other posts and the information propagates.
Identify **cascades** – graphs induced by a time ordered propagation of information [Adamic-Adar ‘05] [SDM ‘07]
Cascade shapes (ordered by decreasing frequency)

- 10 million posts and 350,000 cascades

Cascades are mainly stars (wide and bushy trees)

Interesting relation between the cascade frequency and structure
Methodology:

- Each node of the cascade is a blogpost that belongs to a blog
- For each blog compute the baseline sentiment (over all its posts)
  - Subjectivity: absolute deviation from the baseline
  - Positivity: positive deviation
  - Negativity: negative deviation

Question:

- Does sentiment flow in cascade?
Cascades "heats" up early and then cool off. Subjectivity of the child and the parent are correlated. Sentiment flows!
Advantages:

- Unambiguous, precise and explicit way to trace information flow
- We obtain both the times as well as the trace (graph) of information flow

Caveats:

- Not all links transmit information:
  - Navigational links, templates, adds
- Many links are missing:
  - Mainstream media sites do not create links
  - Bloggers “forget” to link the source
    - (We will later see how to identify networks/cascades just based on what times sites mentioned information)
Complete social media data is near impossible to collect [de Choudhury et al., ‘10]

Missing data and unobserved nodes bias the results

- Estimating influence or a red node gives biased result

Can we correct for such biases?
What happens with missing data?

Data about node $r$ is missing!
Goal: Find properties $X$ of the complete cascade $C$
- We only have access to cascade $C'$ that is $C$ with missing data
  - Each node of $C$ is missing with probability $p$

Results [WSDM ‘11]:
- Our method is most effective when more than 20% of the data is missing
- Works well even with 90% of the data missing
Twitter information network:
- Each user generates a stream of tweets
- Users then subscribe to “follow” the streams of others

3 ways to track information flow in Twitter:
- (1) Trace the spread of a “hashtag” over the network
- (2) Trace the spread of a particular URL
- (3) Re-tweets
Tracing information on Twitter (1)

(1) Tracing hashtags:
- Users annotate tweets with short tags
- Tags naturally emerge from the community
- Given the Twitter network and time stamped posts
  - If user A used hashtag #egypt at $t_1$ and user B follows A and B first used the same hashtag at some later time this means A propagated information to B
(2) Tracing URLs:

- Many tweets contain shortened (hashed) URLs
  - Short-URLs are “personalized”
    - If two users shorten the same URL it will shorten to different strings

- Given the Twitter network and time stamped posts
  - If user A used URL₁ at t₁ and B follows A and B used the same URL later then A propagated information to B
(3) Re-tweets:

- Explicit information diffusion mechanism on Twitter
- B sees A’s tweet and “forwards” it to its follower by re-tweeting
- By following re-tweet cascades we establish the information flow
Meme: A unit of cultural inheritance

Extract textual fragments that travel relatively unchanged, through many articles:

- Look for phrases inside quotes: “…”
  - About 1.25 quotes per document in Spinn3r data

Why it works?

Quotes...

- are integral parts of journalistic practices
- tend to follow iterations of a story as it evolves
- are attributed to individuals and have time and location
Challenge: Quotes Mutate

Quote: Our opponent is someone who sees America, it seems, as being so imperfect, imperfect enough that he's palling around with terrorists who would target their own country.
Goal: Find mutational variants of a quote
Form approximate quote inclusion graph
  - Shorter quote is approximate substring of a longer one
Objective: In DAG (approx. quote inclusion), delete min total edge weight s.t. each connected component has a single “sink”
DAG-partitioning is NP-hard but heuristics are effective:

- **Observation**: enough to know node’s parent to reconstruct optimal solution

- **Heuristic**: Proceed top down and assign a node (keep a single edge) to the strongest cluster
Insights: Quotes reveal pulse of media

http://memetracker.org

Volume over time of top 50 largest total volume quote clusters

August

October

8/21/2011

Jure Leskovec: Social Media Analytics (KDD '11 tutorial)
Can classify individual sources by their typical timing relative to the peak aggregate intensity

<table>
<thead>
<tr>
<th>Rank</th>
<th>Lag [h]</th>
<th>Reported</th>
<th>Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-26.5</td>
<td>42</td>
<td>hotair.com</td>
</tr>
<tr>
<td>2</td>
<td>-23</td>
<td>33</td>
<td>talkingpointsmemo.com</td>
</tr>
<tr>
<td>4</td>
<td>-19.5</td>
<td>56</td>
<td>politicalticker.blogs.cnn.com</td>
</tr>
<tr>
<td>5</td>
<td>-18</td>
<td>73</td>
<td>huffingtonpost.com</td>
</tr>
<tr>
<td>6</td>
<td>-17</td>
<td>49</td>
<td>digg.com</td>
</tr>
<tr>
<td>7</td>
<td>-16</td>
<td>89</td>
<td>breitbart.com</td>
</tr>
<tr>
<td>8</td>
<td>-15</td>
<td>31</td>
<td>thepoliticalcarnival.blogspot.com</td>
</tr>
<tr>
<td>9</td>
<td>-15</td>
<td>32</td>
<td>talkleft.com</td>
</tr>
<tr>
<td>10</td>
<td>-14.5</td>
<td>34</td>
<td>dailykos.com</td>
</tr>
<tr>
<td>30</td>
<td>-11</td>
<td>32</td>
<td>uk.reuters.com</td>
</tr>
<tr>
<td>34</td>
<td>-11</td>
<td>72</td>
<td>cnn.com</td>
</tr>
<tr>
<td>40</td>
<td>-10.5</td>
<td>78</td>
<td>washingtonpost.com</td>
</tr>
<tr>
<td>48</td>
<td>-10</td>
<td>53</td>
<td>online.wsj.com</td>
</tr>
<tr>
<td>49</td>
<td>-10</td>
<td>54</td>
<td>ap.org</td>
</tr>
</tbody>
</table>
Insights: Quotes on Great depression

- Pew’s project for Excellence in journalism
- Media coverage of the current economic crisis

Main

Top republican voice ranks only 14th
There are many other ways to trace information:

- Trace email chain letters [Liben-Nowell-Kleinberg, ‘08]
- Use text classifiers to predict whether there was information flow between two blog posts [Adar-Adamic, ‘05]
- Trace the spread Facebook Page Fans over the Facebook network [Sun et al. ‘09]
- Diffusion of “favoriting” a photo on Flickr [Cha et al. ‘09]
- Product recommendations [Leskovec et al. ‘06]
Part 1: Information flow in networks
  - 1.1: Data collection: How to track the flow?
  - 1.2: Modeling and predicting the flow
  - 1.3: Infer networks of information flow

Part 2: Rich interactions
Q: How does information attention rise and decay? [Wu-Huberman ‘07] [Szabo-Huberman, ‘08]

- **Item i**: Piece of information (e.g., quote, url, hashtag)
- **Volume** $x_i(t)$: # of times $i$ was mentioned at time $t$
  - Volume = number of mentions = attention = popularity
- Q: What are typical classes of shapes of $x_i(t)$?
Given: Volume of an item over time
  i.e., number of mentions of a quote over time

Goal: Want to discover types of shapes of volume time series
Goal: Cluster time series & find cluster centers

Time series distance function needs to be:

\[ d(x, y) = \min_{a, q} \sum_{t} (x(t) - a \cdot y(t - q))^2 \]

- K-Spectral Centroid clustering [WSDM '11]
Patterns of Attention

- **Quotes:** 1 year, 172M docs, 343M quotes
- **Same 6 shapes for Twitter:** 580M tweets, 8M #tags
- **Similar shapes also found in query popularity** [Kulkarni et al. ’11]
Analysis of Attention Patterns

Electric Shock
- Spike created by News Agencies (AP, Reuters)
- Slow & small response of blogs
- Blogs mention 1.3 hours after the mainstream media
- Blog volume = 29.1%

Die Hard
- The only cluster that is dominated by Bloggers both in time and volume
- Blogs mention 20 min before mainstream media
- Blog volume = 53.1%
Different types of media give raise to characteristic popularity/volume patterns
How much attention will information get?
  - How many sites mention information at particular time?

Traditional view:
  - In a network nodes spread information to their neighbors

Problem:
  - The network may be unknown

Idea: Predict the future number of mentions based on who got “infected” in the past
Predicting Information Attention

- How much attention will information get?
  - Who reports the information and when?
    - 1h: Gizmodo, Engadget, Wired
    - 2h: Reuters, Associated Press
    - 3h: New York Times, CNN
  - How many sites will mention the info at time 4, 5, ...?

- Motivating question:
  - If NYT mentions info at time $t$
  - How many additional mentions does this “generate” (on other sites) at time $t+1$, $t+2$, ...?
Idea: Predict the volume based on who got infected in the past

Solution: Linear Influence Model (LIM)

- Assume no network
- Model the global influence of each node
- Predict future volume from node influences

Advantages:

- No knowledge of network needed
- Contagion can “jump” between the nodes
LIM: Strategy

- **K=1 contagion:**
  - \( V(t) \) ... number of new infections at time \( t \)
  - \( M(t) \) ... set of newly infected nodes at time \( t \)

- **How does LIM predict the future number of infections** \( V(t+1) \)?
  - Each node \( u \) has an **influence function**:
    - After node \( u \) gets infected, how many other nodes tend to get infected
    - Estimate the influence function from past data
  - Predict future volume using the influence functions of nodes infected in the past
Node $u$ has an “influence” function $I_u(t)$:

- $I_u(t)$: After node $u$ gets mentions, how many other nodes tend to mention $t$ hours later
  - e.g.: Influence function of CNN: How many sites say the info after CNN says it?
- Estimate the influence function from past data

How to predict future volume $x_i(t+1)$ of info $i$?

- Predict future volume using the influence functions of nodes infected in the past
The Linear Influence Model

LIM model:

- Volume $x_i(t)$ of $i$ at time $t$
- $A_i(t)$ ... a set of nodes that mentioned $i$ before time $t$

And let:

- $I_u(t)$: influence function of $u$
- $t_u$: time when $u$ mentioned $i$

Predict future volume as a sum of influences: 

$$x_i(t + 1) = \sum_{u \in A_i(t)} I_u(t - t_u)$$
Estimating Influence Functions

- After node $u$ mentions the info, $I_u(t)$ other mentions tend to occur $q$ hours later
  - $I_u(t)$ is not observable, need to estimate it
  - We make no assumption about the shape of $I_u(t)$
  - Want to set influence functions $I_u(t)$ such that we minimize the error:

$$\sum \sum \left[ x_i(t+1) - \sum_{u \in A_i(t)} I_u(t-t_u) \right]^2$$
Discrete non-parametric influence functions:

- Discrete time units
- $I_u(t)$ ... non-negative vector of length $L$

$$I_u(t) = [I_u(1), I_u(2), I_u(3), \ldots, I_u(L)]$$

Find $I_u(q)$ by solving a least-squares-like problem:

$$\min_{I_u, \forall u} \sum_i \sum_t \left( x_i(t + 1) - \sum_{u \in A_i(t)} I_u(t - t_u) \right)^2$$
- **Input data:** $K$ contagions, $N$ nodes
- **Write LIM as a matrix equation:**

- **Volume vector:**
  $V_k(t)$ ... volume of contagion $k$ at time $t$

- **Infection indicator matrix:**
  $M_{u,k}(t) = 1$ if node $u$ gets infected by contagion $k$ at time $t$

- **Influence functions:**
  $I_u(t)$ ... influence of node $u$ on diffusion
Estimating influence functions

- **LIM as a matrix equation:** $V = M \times I$
- **Estimate influence functions:**

$$\hat{I} = \arg\min_{I \geq 0} \|V - M \cdot I\|^2_2$$

- **Solve using Non-Negative Least Squares**
  - Well known, can use Reflective Newton Method
  - Time $\sim 1$ sec when $M$ is 200,000 x 4,000 matrix
- **Predicting future volume:** **Simple!**
  - Given $M$ and $I$, then
  - $V = M \times I$
Take top 1,000 quotes by the total volume:
- Total 372,000 mentions on 16,000 websites
- Build LIM on 100 highest-volume websites
  - $x_i(t)$ ... number of mentions across 16,000 websites
  - $A_i(t)$ ... which of 100 sites mentioned quote $i$ and when
- Improvement in L2-norm over 1-time lag predictor

<table>
<thead>
<tr>
<th></th>
<th>Bursty phrases</th>
<th>Steady phrases</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>7.21%</td>
<td>8.30%</td>
<td>7.41%</td>
</tr>
<tr>
<td>ARMA</td>
<td>6.85%</td>
<td>8.71%</td>
<td>7.75%</td>
</tr>
<tr>
<td>LIM (N=100)</td>
<td>20.06%</td>
<td>6.24%</td>
<td>14.31%</td>
</tr>
</tbody>
</table>
Influence functions give insights:

Q: NYT writes a post on politics, how many people tend to mention it next day?
A: Influence function of NYT for political phrases!

Experimental setup:

5 media types:
- Newspapers, Pro Blogs, TVs, News agencies, Blogs

6 topics:
- Politics, nation, entertainment, business, technology, sports

For all phrases in the topic, estimate average influence function by media type
Politics is dominated by traditional media

- Blogs:
  - Influential for Entertainment phrases
  - Influence lasts longer than for other media types
Part 1: Information flow in networks

1.1: Data collection: How to track the flow?
1.2: Modeling and predicting the flow
1.3: Infer networks of information flow

Part 2: Rich interactions
But how does information really spread?

We only see time of mention but not the edges

Can we reconstruct (hidden) diffusion network?
There is a **hidden** diffusion network:

We only see **times** when nodes get “infected”:

- $c_1: (a,1), (c,2), (b,3), (e,4)$
- $c_2: (c,1), (a,4), (b,5), (d,6)$

Want to infer who-infests-whom network!
Virus propagation

Viruses propagate through the network

We only observe when people get sick

But NOT who infected whom

Word of mouth & Viral marketing

Recommendations and influence propagate

We only observe when people buy products

But NOT who influenced whom

Can we infer the underlying network?
The optimization problem

- **Goal**: Find a graph $G$ that best explains the observed information times:
  - Given a graph $G$, define the likelihood $P(C|G)$:
    - Define a model of information diffusion over a graph
    - $P_c(u,v)$ ... prob. that $u$ infects $v$ in cascade $c$
    - $P(c|T)$ ... prob. that $c$ spread in particular pattern $T$
    - $P(c|G)$ ... prob. that cascade $c$ occurred in $G$
    - $P(G|C)$ ... prob. that a set of cascades $C$ occurred in $G$
  - **Questions**:
    - How to efficiently compute $P(G|C)$? (given a single $G$)
    - How to efficiently find $G^*$ that maximizes $P(G|C)$? (over $O(2^{N*N})$ graphs)
Consider 1 cascade: the model

- Cascade reaches node \( i \) at time \( t_i \), and spreads to \( i \)'s neighbors \( j \):
  - With prob. \( \beta \) cascade propagates along edge \((u,v)\) and \( t_v = t_u + \Delta \)

Transmission probability:

\[
P_c(u,v) \propto P(t_v - t_u) \quad \text{if} \quad t_v > t_u \quad \text{else} \quad \epsilon
\]

- \( \epsilon \) captures influence external to the network
  - At any time a node can get infected from outside with small probability \( \epsilon \)
Information Diffusion Model

- Given node infection times and pattern $T$:
  - $c = \{ (a,1), (c,2), (b,3), (e,4) \}$
  - $T = \{ a \rightarrow b, a \rightarrow c, b \rightarrow e \}$

- Prob. that $c$ propagates in pattern $T$

\[
P(c|T) = \prod_{(u,v) \in E_T} \beta P_c(u,v) \prod_{u \in V_T, (u,x) \in E \setminus E_T} (1 - \beta)
\]

- Edges that “propagated”
- Edges that failed to “propagate”

- Approximate it as:

\[
P(c|T) \approx \prod_{(u,v) \in E_T} P_c(v,u)
\]
How likely is c to spread in graph G?

\[ c = \{(a,1), (c,2), (b,3), (e,4)\} \]

Need to consider all possible ways for c to spread in G (i.e., all spanning trees T):

\[
P(c|G) = \sum_{T \in \mathcal{T}_c(G)} P(c|T) \approx \max_{T \in \mathcal{T}_c(G)} P(c|T)
\]

Consider the most likely propagation tree
Optimization problem

- Score of a graph $G$ for a set of cascades $C$:
  \[
  P(C|G) = \prod P(c|G) \\
  F_C(G) = \sum_{c \in C} \log P(c|G)
  \]

- Want to find the “best” graph:
  \[
  G^* = \arg\max_{|G| \leq k} F_C(G)
  \]

The problem is NP-hard:
MAX-k-COVER [KDD ’10]
Theorem: Function $F_C(G)$ is monotonic, and submodular in edges of $G$:

- Let $A$, $B$ be two graphs: same nodes, different edges: $A \subseteq B \subseteq V \times V$:

\[
F_C(A \cup \{e\}) - F_C(A) \geq F_C(B \cup \{e\}) - F_C(B)
\]

Gain of adding an edge to a “small” graph \hspace{1cm} Gain of adding an edge to a “large” graph

Benefits:

1. Efficient (and simple) optimization algorithm
2. Approximation guarantee ($\approx 0.63$ of OPT)
3. Tight on-line bounds on the solution quality
NetInf: The Algorithm

- **NetInf algorithm:**
  Use greedy hill-climbing to maximize $F_C(G)$:
  - Start with empty $G_0$ (G with no edges)
  - Add $k$ edges ($k$ is parameter)
  - At every step add an edge to $G_i$ that maximizes the marginal improvement

\[
e_i = \arg\max_{e \in G \setminus G_{i-1}} F_C(G_{i-1} \cup \{e\}) - F_C(G_{i-1})
\]
Experiments: Synthetic data

- **Synthetic data:**
  - Take a graph $G$ on $k$ edges
  - Simulate info. diffusion
  - Record node infection times
  - Reconstruct $G$

- **Evaluation:**
  - How many edges of $G$ can NetInf find?
    - Break-even point: 0.95
    - Performance is independent of the structure of $G$!
Example: Real Data

- Memetracker quotes:
  - 172 million news and blog articles
  - Aug '08 – Sept '09
  - Extract 343 million phrases
  - Record times $t_i(w)$ when site $w$ mentions quote $i$

- Given times when sites mention quotes
- Infer the network of information diffusion:
  - Who tends to copy (repeat after) whom
Example: Diffusion Network

- 5,000 news sites:

Blogs
Mainstream media
Diffusion Network (small part)
Detecting information outbreaks

Detect **blue** & **yellow** soon but miss **red**.

Want to read things **before** others do.

Detect **all** stories but **late**.
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 know the story before we do
Optimization problem

- **Given:**
  - Graph G(V,E), budget C
  - Data on how cascades spread over time

- **Select a set of nodes A maximizing the reward**
  
  \[
  \max_{A \subseteq V} \sum_{i} \text{Prob}(i) R_i(A)
  \]
  
  subject to \( \text{cost}(A) \leq C \)

- **Solving the problem exactly is NP-hard**
  - Set cover [Kuhler et al. ’99]
Problem structure: Submodularity

- Gain of adding a node to small set is larger than gain of adding a node to large set
- **Submodularity**: diminishing returns
- **Algorithm**: 
  - Greedily add node that gives highest increase in reward
We must show $R$ is submodular: $A \subseteq B$

$$R(A \cup \{u\}) - R(A) \geq R(B \cup \{u\}) - R(B)$$

Gain of adding a node to a small set

Gain of adding a node to a large set

Natural example:

- Sets $A_1, A_2, \ldots, A_n$
- $R(A) =$ size of union of $A_i$
  (size of covered area)

If $R_1, \ldots, R_K$ are submodular, then $\sum R_i$ is submodular
Theorem:
- Reward function is submodular

Consider cascade $i$:
- $R_i(u_k) =$ set of nodes saved from $u_k$
- $R_i(A) =$ size of union $R_i(u_k), u_k \in A$

$\Rightarrow R_i$ is submodular

Global optimization:
- $R(A) = \sum R_i(A)$

$\Rightarrow R$ is submodular
We develop CELF 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: $\text{arg max}(R(A'), R(A''))$

**Theorem:** If $R$ is submodular then CELF near optimal:

CELF achieves $\frac{1}{2}(1-1/e)$ factor approximation
Given a budget (e.g., of 3 blogs)
Select sites to cover the most of the network
Question: Which websites should one read to catch big stories?

Idea: Each blog covers part of the network.

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

Which blogs to read to be most up to date?

<table>
<thead>
<tr>
<th>Number of selected blogs</th>
<th>Our solution</th>
<th>In-links (used by Technorati)</th>
<th>Out-links</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

% of stories detected (higher is better)
Messages arriving through networks from real-time sources requires new ways of thinking about information dynamics and consumption:

- Tracking information through (implicit) networks
- Quantify the dynamics of online media
- Predict the diffusion of information
- And infer networks of information diffusion
Can this analysis help identify dynamics of polarization [Adamic-Glance ‘05]?

Connections to mutation of information:

- How does attitude and sentiment change in different parts of the network?
- How does information change in different parts of the network?
References

[Adar-Adamic ‘05] E. Adar and L. A. Adamic, Tracking Information Epidemics in Blogspace, Web Intelligence, 2005


