BUZZTRAQ

Predicting geographical access patterns of social cascades using social networks

Nishanth Sastry
Eiko Yoneki & Jon Crowcroft
University of Cambridge
LONG TAIL

- "Life in the vast lane." – Jim Treacher
- "An embarrassment of niches." – Kevin Kelleher
- "We sold more books today that didn't sell at all yesterday than we sold today of all the books that did sell yesterday." – Former Amazon employee Josh Petersen

THE DYNAMIC VERSION OF LONG-TAIL

usually depicted as a static picture - few objects enormously popular, large number of objects, which are each insignificant but collectively popular
THE DYNAMIC VERSION OF LONG-TAIL
THE DYNAMIC VERSION OF LONG-TAIL

- Users need to become aware of new rare objects
- A delivery infrastructure is needed

With the proper infrastructure, a rare object can become popular very quickly! -- The "Lily Allen Effect"
DISTRIBUTING RICH-MEDIA LONG-TAIL CONTENT

- Rich media content has strict delivery constraints
- Delivery needs custom-built infrastructure
  - e.g. caching and replica placement close to user

- BUT: content objects in the tail are accessed rarely
  - not worth akamaizing any given tail object
  - collectively, worth replicating!

Is there any way out? The key to the solution lies in how the objects are accessed.
PATTERNS OF ACCESS

<table>
<thead>
<tr>
<th>Means of awareness</th>
<th>Delivery infrastructure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Popular content</td>
<td>Broadcast highlights</td>
</tr>
<tr>
<td>Long-tail content</td>
<td>CDN</td>
</tr>
</tbody>
</table>
# Patterns of Access

<table>
<thead>
<tr>
<th>Popular content</th>
<th>Means of awareness</th>
<th>Delivery infrastructure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Broadcast highlights</td>
<td>CDN</td>
</tr>
<tr>
<td>Long-tail content</td>
<td>Social cascade (Cha ’06)</td>
<td>Selective replica placement</td>
</tr>
</tbody>
</table>

**How to place replicas selectively?**
PROBLEM SETUP

- How to deliver content not (yet) globally popular?
- Servers exist in multiple geographical regions
- An object can be replicated on $k$ of these

Given an object, how do we predict the best servers/regions?
REPLICA PLACEMENT STRATEGIES

- **Social cascade prediction (Buzztraq):** Map geographic regions of user's friends. Place replicas in the $k$ regions with most friends.
- **Location based placement:** Place replicas in the $k$ regions with most number of previous users.
- Look up user geographies from their IP address.

Does social cascade prediction offer benefits? Depends on the nature of the social cascade.
In this case, social cascade prediction gives the same prediction as location-based placement, so it will be at least as good.
SHIFTING SOCIAL CASCADE

The new users were friends of previous users. Social cascade prediction counts them with the previous user’s visit and predicts the new regions that gain or lose accesses.

Next, how to implement SCP?
BUZZTRAQ PROTOTYPE

- Obtains friends of each user, and their affiliations, from Facebook
- Google geocoding API is used on affiliations in an attempt to deduce geographic affiliations
  - Mostly successful: latitude-longitude co-ordinates obtained for 71.3% of 1660 affiliations evaluated
- Predict next access from regions with most friends
POTENTIAL PITFALLS

- Geocoding sometimes inadequate/inaccurate
- Using union of geographical affiliations may lead to inaccurate social cascade predictions

- These can be compensated by aggregating into regions
  - If all affiliations of a user belong to same region, Buzztraq hints will not be wrong!
  - For users more likely to spread the cascade in regions where they have more friends, Buzztraq hints will be correct.
HOW USEFUL IS BUZZTRAQ?

- Expected utility depends on collective weight of tail
- The least popular 90% of videos
  - = 40% of accesses (Yu’06)
  - = 20% of accesses (Cha’07)
EVALUATION - USERS

- Users accesses drawn from 20,740 facebook users
  (Harvard network, IDs < 36,000, 1181 locations)

- Social cascade component in user access sequence
  - With probability $p_s$, next user is friend of previous
  - With probability $1-p_s$, next user is randomly chosen
EVALUATION - REGIONS

- World divided into 10 regions by k-means clustering; 4 in US, which has majority of users
- Content Provider allowed 3 replicas
  - forced to serve at least one US region remotely
EVALUATION - HOW MANY LOCAL ACCESSES?

- Design cost function
- Make remote access costlier than local access
- Measure cumulative cost of serving first \( n \) requests
SOCIAL CASCADE PREDICTION IS CHEAPER

![Graph showing cost ratio vs number of accesses]
...WHETHER SOCIAL CASCADE IS A MAJOR OR MINOR FACTOR

- \( p = 0 \) is random access, every strategy just as good.
- \( p = 1 \) is pure social cascade. Heatmap from before.
- \( p = 0.5 \) has been studied in previous slide.
- Other vals here.
STILL EARLY WORK...

- More sophisticated models for social cascade
- This work answers where to place replicas of a given content object. Will have to choose which to replicate when bandwidth or storage is constraint
- Answer is not straightforward popularity based: confidence of the social cascade prediction may be higher for other (less popular) objects
BUZZTRAQ

- Estimate **future** users from friends of current users
- Use that to predict where to place replicas of content