(Un)Predictability of Social Networks

Lei Tang
References

- *Experimental Study of Inequality & Unpredictability in an Artificial Cultural Market*, Science, 2006
- *Prediction of Popularity of Digg & Youtube*
- *Link Prediction Problem in Social Network*, 2005
- *The Black Swan: The Impact of the Highly Improbable*
Hit songs, books and movies are many times more successful than average, suggesting that "the best" alternatives are qualitatively different from "the best"; yet experts routinely fail to predict which products will succeed.

- Black Swan Effect?
- What for predict?
Two Views

- Inequality & Unpredictability
  How can success in cultural markets be strikingly distinct from average performance and yet so hard to anticipate?

- Quality Model
  - mapping from "quality" to success is convex.
  - Cannot explain unpredictability.

- Influence Model
  - Individuals do not make decisions independently.
  - Collective decisions with social influence exhibits extreme variation.

- Empirical Verification is missing.
Challenges

- Requires comparisons of multiple realization of stochastic process
  - Parallel Universe
- In reality, only one "history" is observed.
  - History is not repeatable.
- Design an experiment with online service to study social influence in cultural market.
Experiment Setup

- An artificial "music market"
  - 14,341 participants
  - 48 songs from 18 unknown bands
  - Users are randomly assigned to a "universe"

- Users
  - Listen to the song
  - Assign a rating
  - Opportunity to download the song.
Different Experimental Conditions

<table>
<thead>
<tr>
<th>Layout</th>
<th>Independent</th>
<th>Social Influence</th>
</tr>
</thead>
<tbody>
<tr>
<td>16X3 rectangular grid, with positions of songs randomly assigned.</td>
<td>Exp1-independent</td>
<td>Exp1-Social Influence</td>
</tr>
<tr>
<td>One column of songs sorted by download count</td>
<td>Exp2-independent</td>
<td>Exp2-Social Influence</td>
</tr>
</tbody>
</table>

For Social Influence, 8 independent "universe" were studied.
Inequality (diff among different songs)

\[ G = \sum_{i=1}^{S} \sum_{j=1}^{S} |m_i - m_j| / 2S \sum_{k=1}^{S} m_k \]

\[ 0 \leq G \leq 1 \]
Unpredictability (diff of different worlds)

\[ u_i = \sum_{j=1}^{W} \sum_{k=j+1}^{W} |m_{i,j} - m_{i,k}| / \left(\binom{W}{2}\right) \]

![Graph A: Experiment 1](image)

![Graph B: Experiment 2](image)
Relationship between Quality & Success

A

Exp. 1

$m_{\text{influence}}$ vs $m_{\text{indep}}$

B

Exp. 1

Rank: $m_{\text{influence}}$

C

Exp. 2

$m_{\text{influence}}$ vs $m_{\text{indep}}$

D

Exp. 2

Rank: $m_{\text{influence}}$
Relationship between Quality & Success

- the "best" songs never do very badly, and the "worst" songs never do extremely well.
- The "best" songs are most unpredictable.
- The larger the social influence is, the unpredictable it is.
Ranks of Songs in Different Worlds
Conclusions & Further Questions

- Limitations: more solid to have multiple replica of independent worlds.
- Social Influence leads to extreme variance.
- Quality alone is incomplete for prediction.
- So a conservative question is:
  - Could we infer the "success" from early stage of the social influence?
Predicting the Popularity

- **YouTube**
  - collect view count time series on 7,146 selected videos daily
  - Beginning from Apr. 21th, 2008
  - Videos are collected from "recently added" to avoid bias

- **Digg**
  - Retrieve all diggs made by registered users between 07/01/2007 - 12/18/2007
  - 60 million diggs, 850,000 users, 2.7 million submissions
Bias of Digging activity

weekends

midnight
Activity Granularity

- The average number of diggs arriving to promoted stories per hour is 5,478.
- One digg hour: the time it takes for so many new diggs to be cast.
- For YouTube, focus on daily as youtube update the count no more than once every day.
Correlation

Digg

You Tube

Strong Linear Correlation
Strong Linear Correlation

![Graph showing Pearson correlation coefficient over time for different data sets. The graph includes lines for YouTube (days), YouTube (untr.), Digg (digg hours), and Digg (hours). The x-axis represents time in days, hours, and Digg hours, while the y-axis represents the Pearson correlation coefficient, ranging from 0.7 to 1.0.]
Prediction

- Linear regression on a logarithmic scale (LN)
  - least-squares absolute error
- Constant Scaling Model (CS)
  - Relative squared error
    \[ RSE = \sum_c \left( \frac{\hat{N}_c(t_i, t_f) - N_c(t_f)}{N_c(t_f)} \right)^2 \]
- Growth Profile Model (GP)
  - Assume the mean of popularity grows linearly
Predictive Performance

(a) Squared error vs. Digg story age (digg days)
(b) Relative squared error vs. Digg story age (digg days)
(c) Squared error vs. Youtube video age (days)
(d) Relative squared error vs. Youtube video age (days)
Difference between Digg & Youtube
Comments

- The popularity of content can be predicted very soon after the submission has been made based on early-stage popularity.
- Due to the large variance, relative squared error is more reasonable to estimate the prediction.
- Two possible applications:
  - advertising (more on relative error)
  - content ranking (more on absolute error, difficult)
Other prediction problems

- Link Prediction
  - Whether two actors will be connected at certain time stamp

- Existing Approaches
  - Unsupervised:
    - use various similarity measure
  - Supervised:
    - extract structural features to learn a mapping function

- Performance: Far from satisfactory
  - e.g. accuracy, random (0.15% - 0.48%)
  - using similarity, increase by a facor of 50%
  - still low!
Discussions

- Social Network is highly dynamic
- With collective influence, the outcome is difficult to predict.
- With early stage popularity, it is possible to esitamte the popularity at later stage.
- Accurate link prediction remains a challenge.
- Can we predict more on social network?