(Un)Predictabilty of Social Networks

Lei Tang

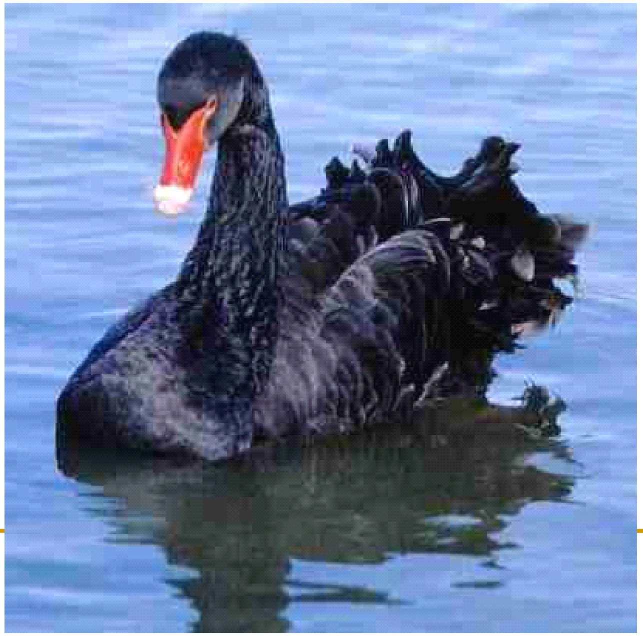
References

- *Experimental Study of Inequality & Unpredictability in an Artifical 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*

Predictability

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 strinkingly 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 influnce exhibits extreme variation.
- Empirical Verification is missing.

Challenges

- Requires comparisions of multiple realization of stochastic process
  - Parallel Universe
- In reality, only one "history" is observed.
  - History is not repeatble.
- 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 unkown bands
  - Users are randomlly assign to a "universe"
- Users
  - listen to the song
  - assign a rating
  - opportunity to download the song.

Different Experimental Conditions

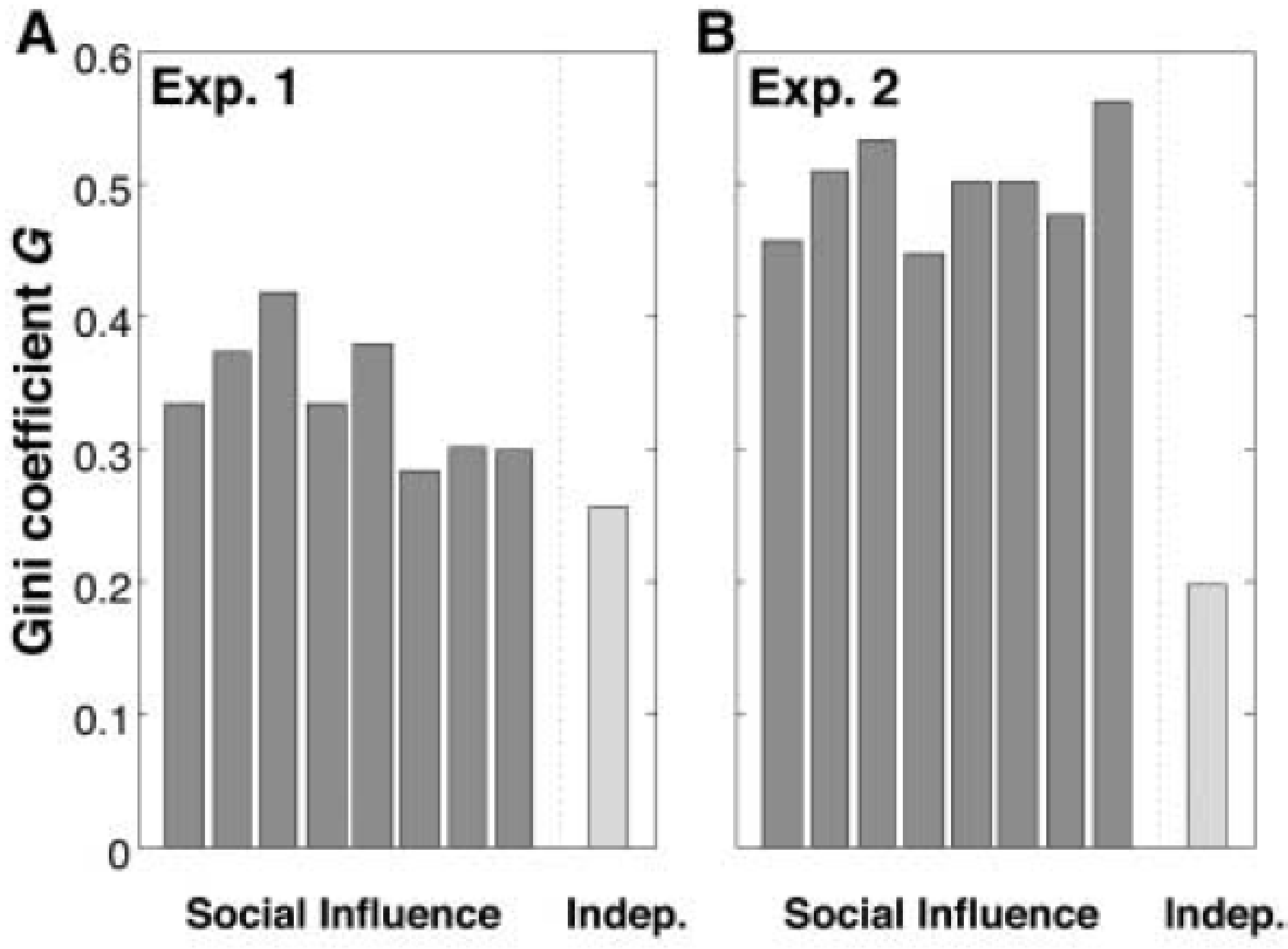
Layout	Independent Names only; No preference information of others	Social Influence Preference information of others included.
16X3 rectangular grid, with positions of songs randomly assigned.	Exp1-independent	Exp1-Social Influence
One column of songs sorted by download count	Exp2-independent	Exp2-Social Influence

For Social Influence, 8 indpendent "universe" were studied.

Inequality (diff among different songs)

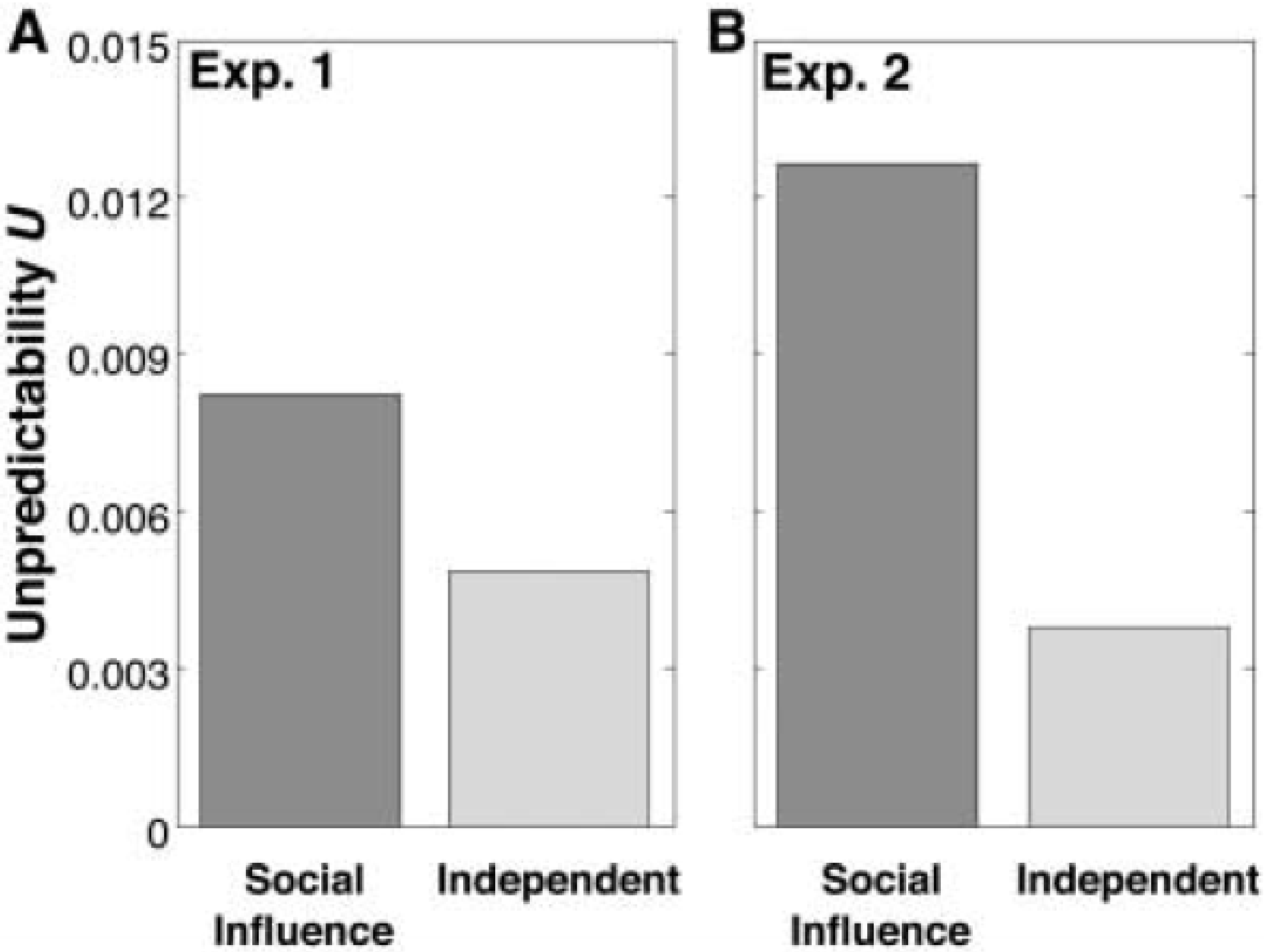
$$G = \sum_{i=1}^{S} \sum_{j=1}^{S} |m_i - m_j| / 2S \sum_{k=1}^{S} m_k$$

0<=G<=1

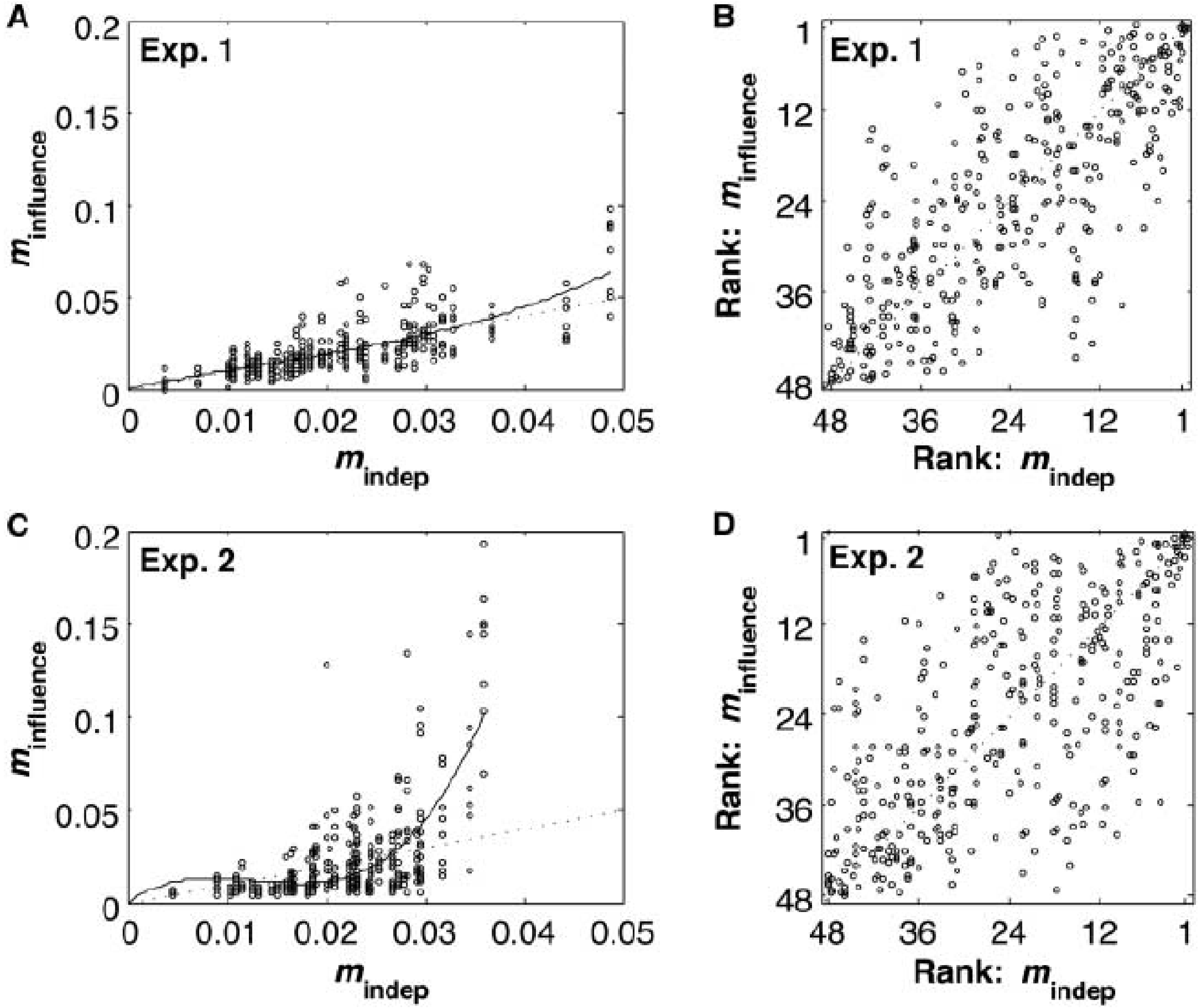


Unpredictability (diff of different worlds)

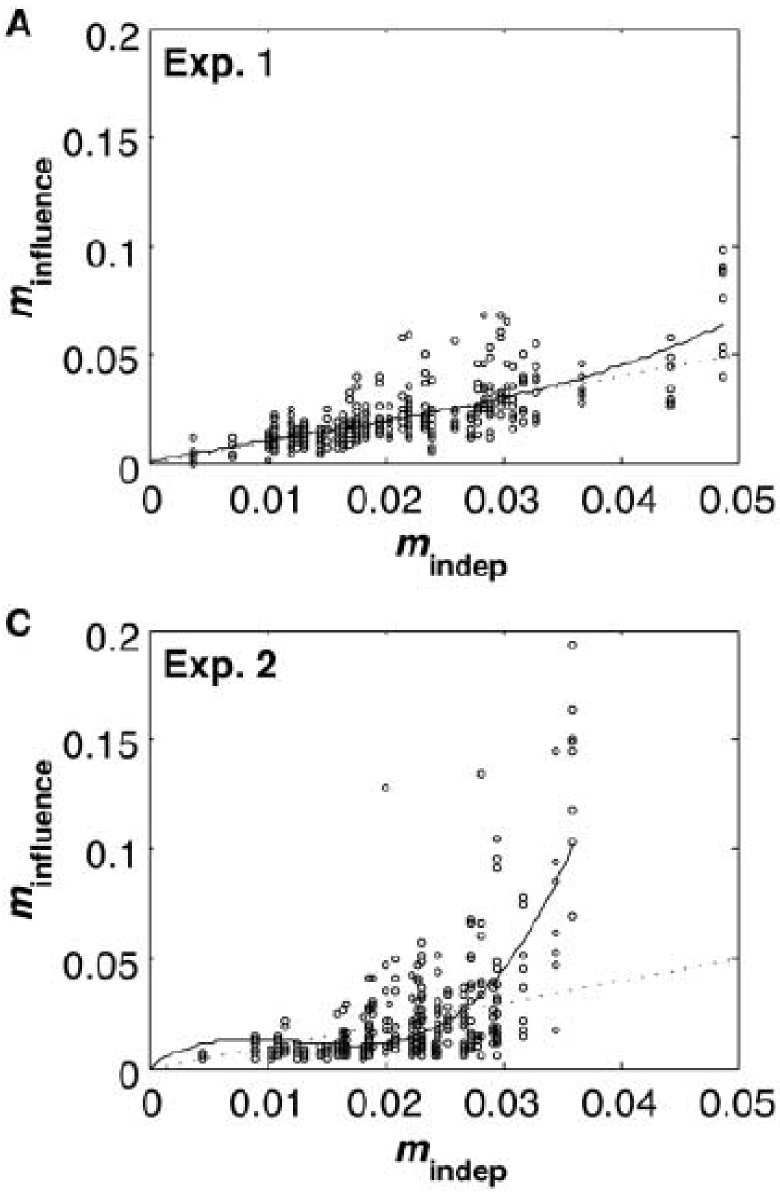
$$u_i = \sum_{j=1}^{W} \sum_{k=j+1}^{W} |m_{i,j} - m_{i,k}| / \binom{W}{2}$$



Relationship between Quality & Success

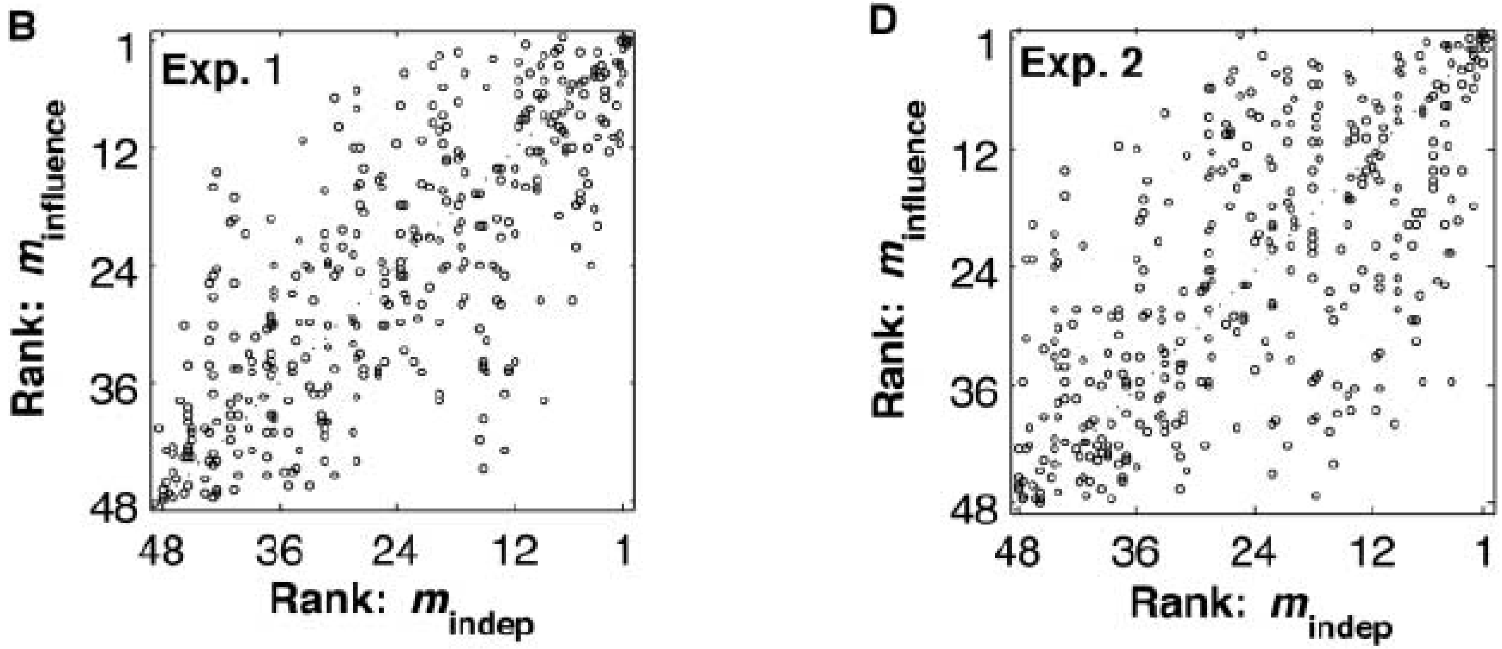


Relationship between Quality & Success



- 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 unpreditable it is.

Ranks of Songs in Different Worlds



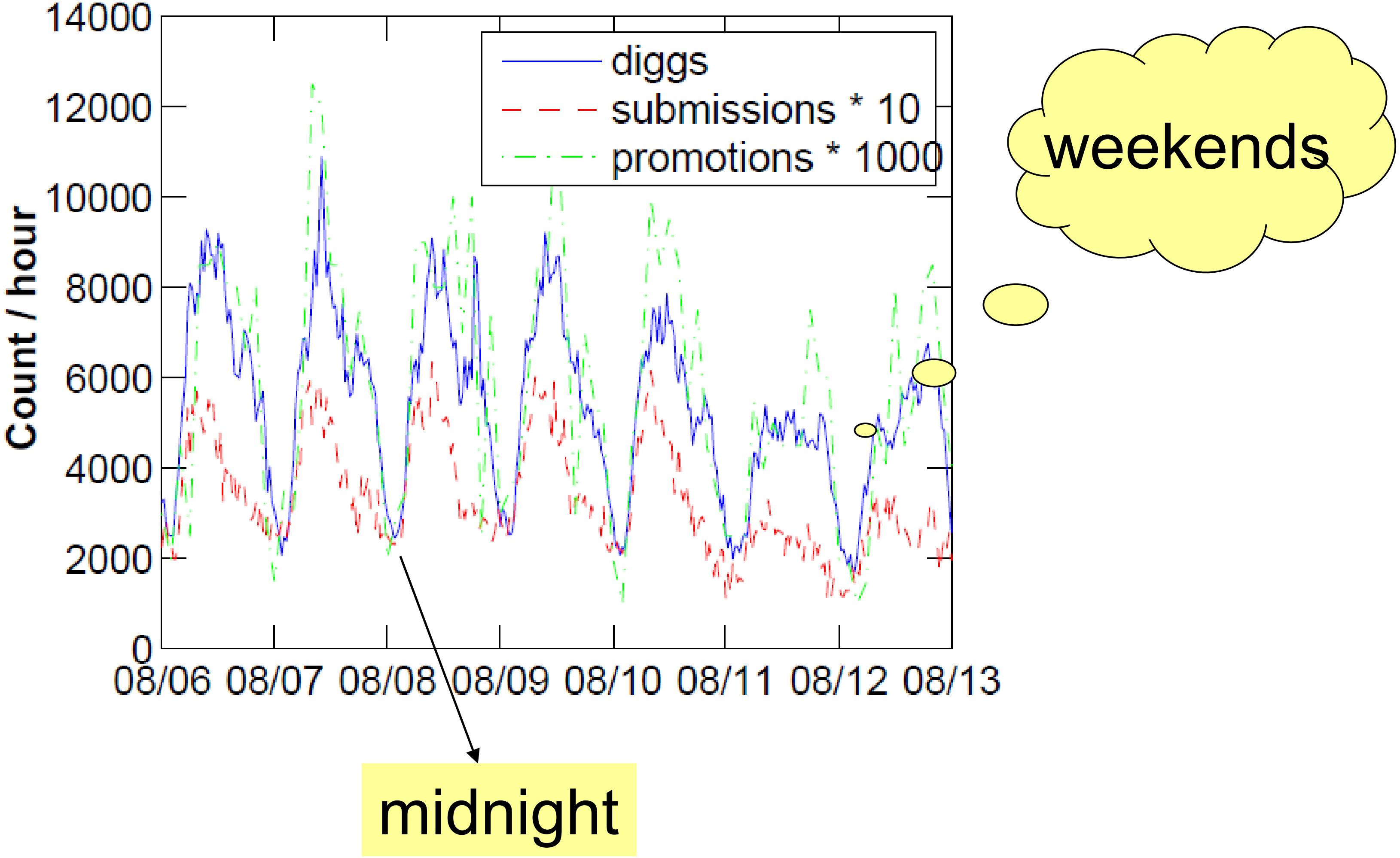
Conclusions & Furthur Questions

- 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 slected videos daily
  - Begining 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

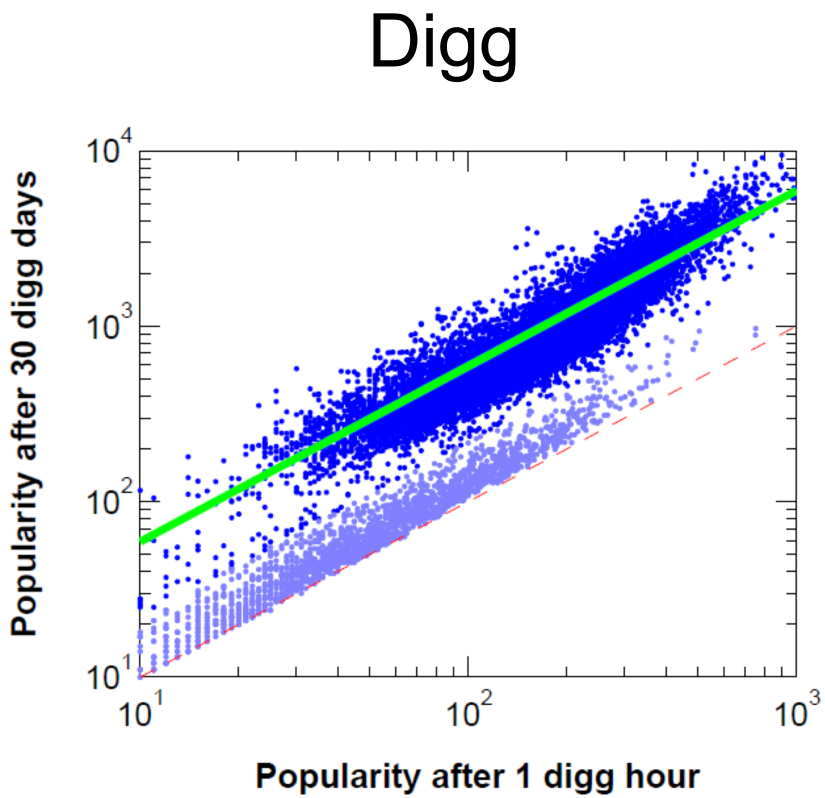
Bias of Digging activity

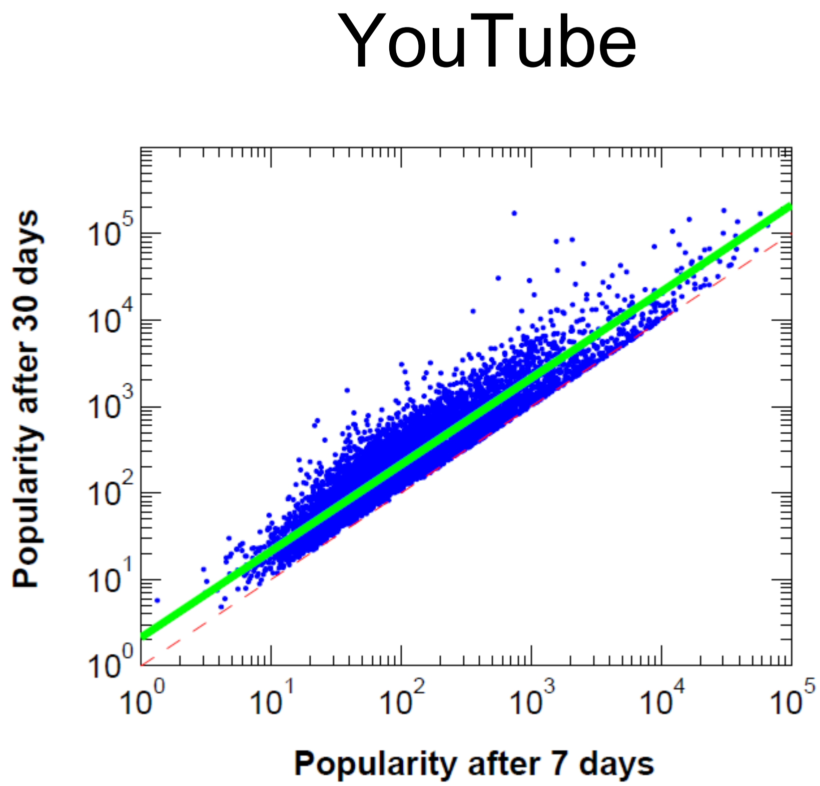


Activity Granularity

- 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 eady day.

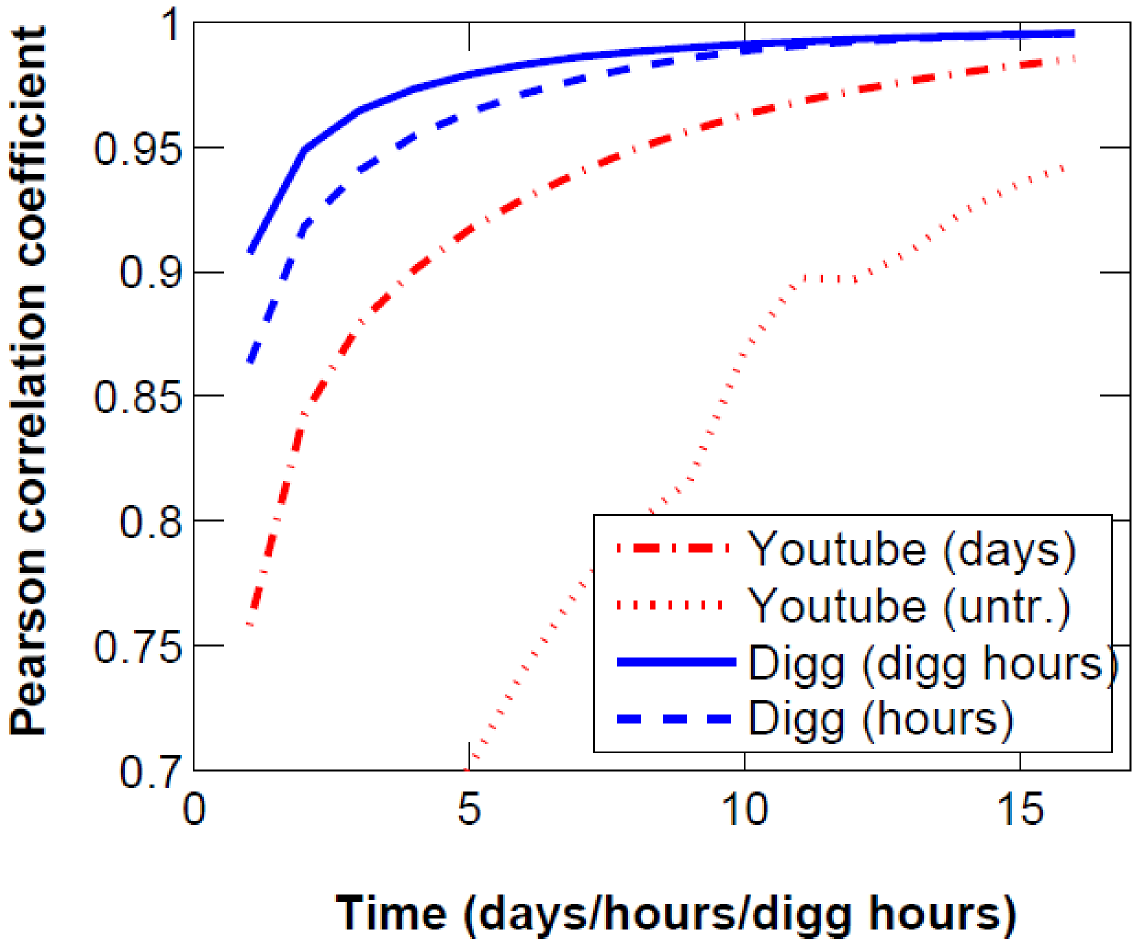
Correlation





Strong Linear Correlation

Strong Linear Correlation



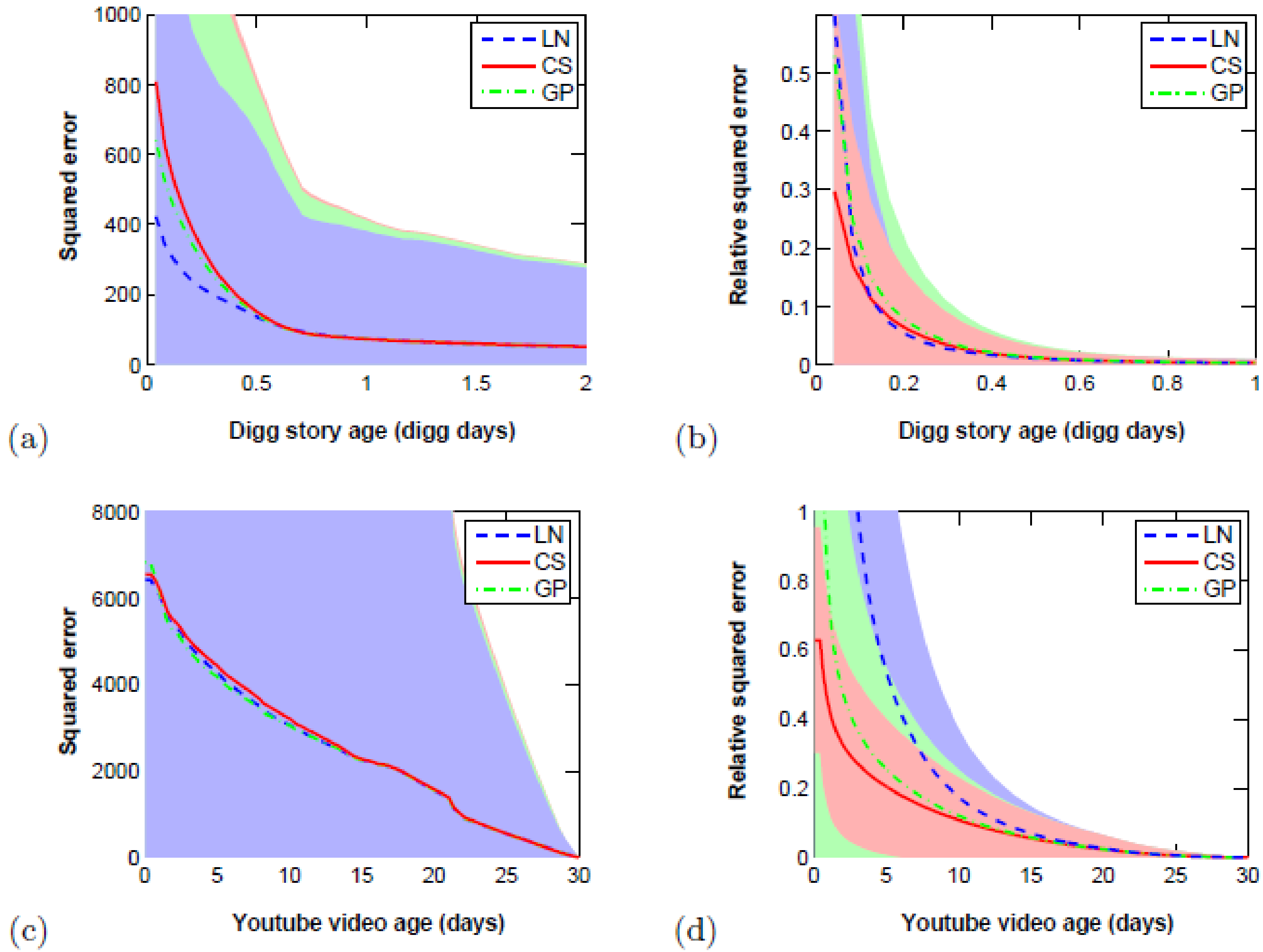
Prediction

- Linear regression on a logarithmic scale (LN)
  - least-squares absolute error
- Constant Scaling Model (CS)
  - Relative squared error

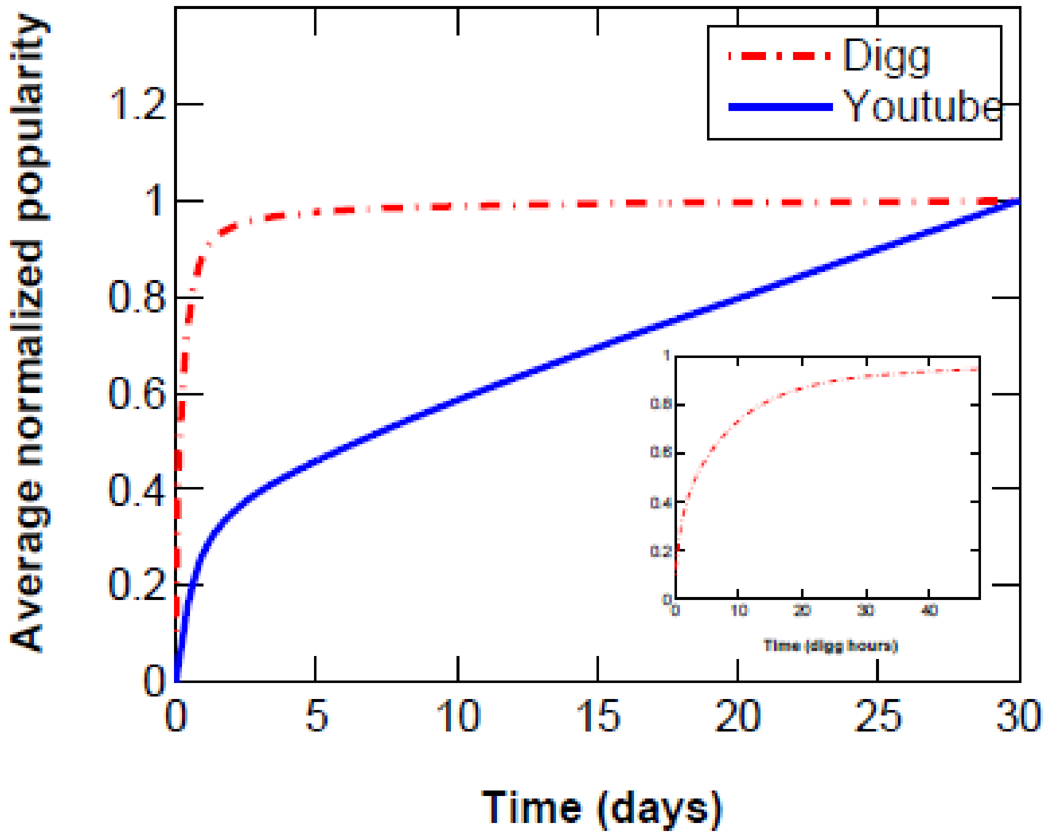
$$\mathrm{RSE} = \sum_{c} \left[\frac{\hat{N}_c(t_i, t_r) - N_c(t_r)}{N_c(t_r)} \right]^2$$

- Growth Profile Model (GP)
  - Assume the mean of popularity grows linearly

Predictive Performance



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 Netowork 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?