Influence and Correlation in Social Networks

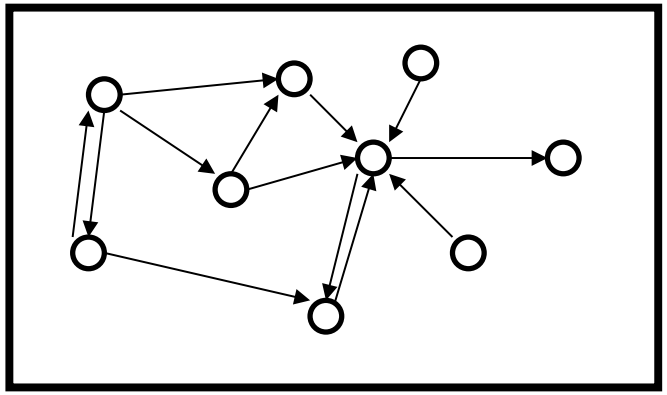
Aris Anagnostopoulos
Department of Informatics and System Sciences
Sapienza University of Rome

Based on joint work with

Ravi Kumar and **Mohammad Mahdian**
Yahoo! Research

Social Systems

- Social network: graph that represents relationships between independent agents.



- Offline social networks
  - Friendships
  - Professional contacts
- Online social networks
  - Facebook, myspace
  - Content-sharing systems: Flickr, YouTube
  - Content-creation systems: Wikipedia

Social Correlation

- How similar is the behavior of connected agents.

- Examples of studies:
  - Offline behavior
    - Fashion
    - Happiness [Fowler and Christakis]
  - Online behavior
    - Joining online communities [Backstrom et al.]
    - Tagging vocabulary on Flickr [Marlow et al.]
    - Use a premium Voice-over-IP service

flickr LOVES YOU™

You aren't signed in Sign In Help

Home The Tour Sign Up Explore Search

piazza san marco

ALL SIZES

piazza san marco, venice

This photo has notes. Move your mouse over the photo to see them.

Comments

mac on a mac pro says:

Wonderful!

Posted 7 months ago. (permalink)

~~ Reza ~~ pro says:

A nice action shot!

Posted 7 months ago. (permalink)

Uploaded on November 23, 2007
by mmahdian

mmahdian's photostream

94 uploads

browse

This photo also belongs to:

faves (Set)

17 items

browse

Tags

- venice
- venezia
- italy
- italia
- st mark square
- piazza san marco
- birds
- girl

Additional Information

Sources of Correlation

- Social influence (induction)

  One person performing an action can cause her contacts to do the same.
  - by providing information
  - by increasing the value of the action to them
- Homophily (selection):

  Similar individuals are more likely to become friends.
  - Example: two mathematicians are more likely to become friends.
- Confounding factors

  External influence from elements in the environment.
  - Example: A family ate in the same restaurant and got stomach pain.

Social Influence

- Focus on a particular "action" A, e.g.:
  - Buy a product
  - Join an online community
  - Use a particular tag in Flickr
- An agent who performs A is called "active".
- x has influence over y if x performing A increases the likelihood that y performs A.
- Distinguishing factor: causality relationship

Identifying Social Influence

- Why is it important?
- Analysis: predicting the dynamics of the system. Whether a new norm of behavior, technology, or idea can diffuse like an epidemic
- Design: designing a system to induce a particular behavior, e.g.:
  - Vaccination strategies (random, targeting a demographic group, random acquaintances, etc.)
  - Design of health policies (e.g., against smoking)
  - Viral marketing campaigns

Approach

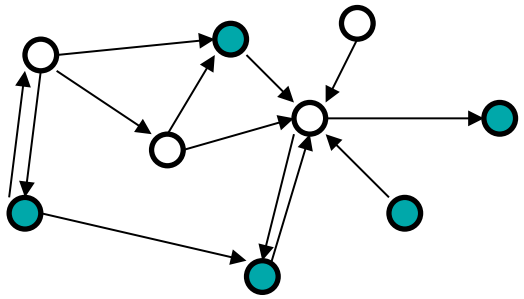
- Measure correlation
- Models for influence and correlation
- Tests for distinguishing influence from correlation
- Theoretical results
- Apply tests on synthetic data
- Apply tests on real data from Flickr

Influence Model

- Graph (static or dynamic)
- Edge (u,v): Node u can influence node v
- Discrete time: t = 0, 1, 2, …
- For each t, every inactive node becomes active with probability p(x), where x is the # active contacts

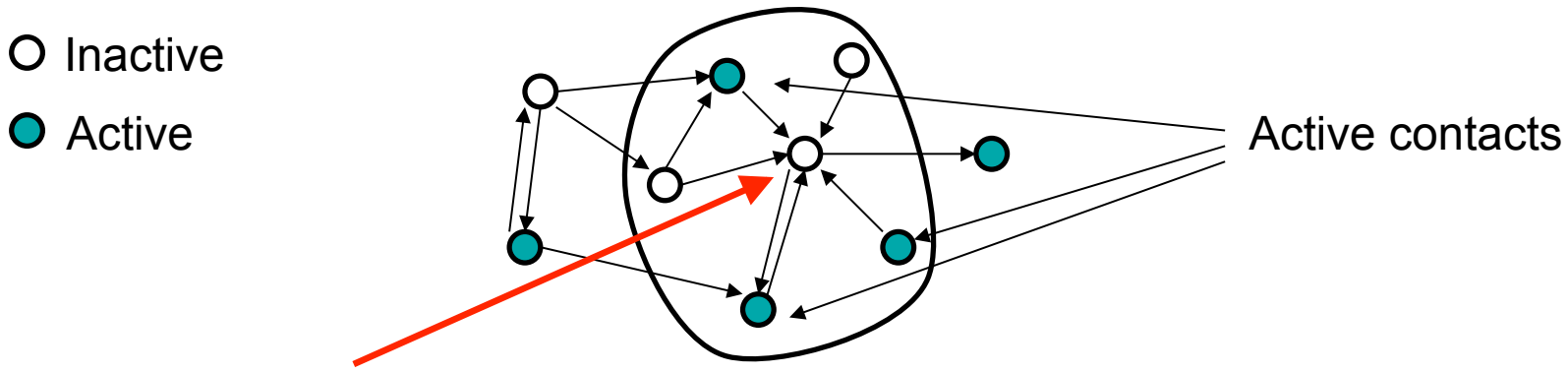
Inactive

Active



Influence Model

- Graph (static or dynamic)
- Edge (u,v): Node u can influence node v
- Discrete time: t = 0, 1, 2, …
- For each t, every inactive node becomes active with probability $p(x)$, where x is the # active contacts



Model – Influence Probability

- Natural choice for $p(x)$: logistic regression function:

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \alpha \cdot x + \beta$$

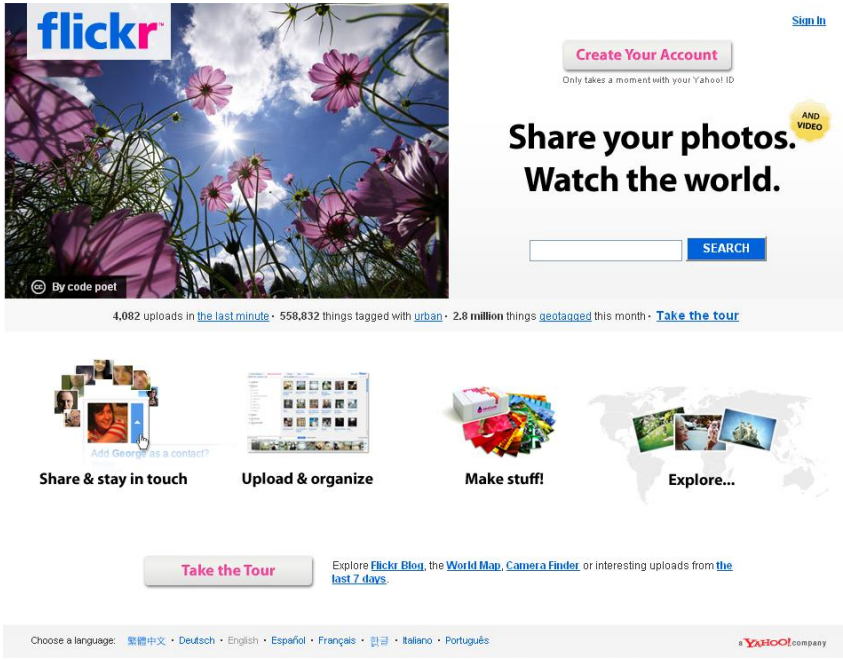
with x (# active contacts) is the explanatory variable. I.e.,

$$p(x) = \frac{e^{\alpha \cdot x + \beta}}{1 + e^{\alpha \cdot x + \beta}}$$

- Given data, can estimate ® with Maximum Likelihood
- Coefficient ® measures social correlation.

Flickr Data Set

- Photo sharing website
- 16 month period
- Growing # of users, final number ~800K
- ~340K users who have used the tagging feature
- Social network:
  - Users can specify “contacts”

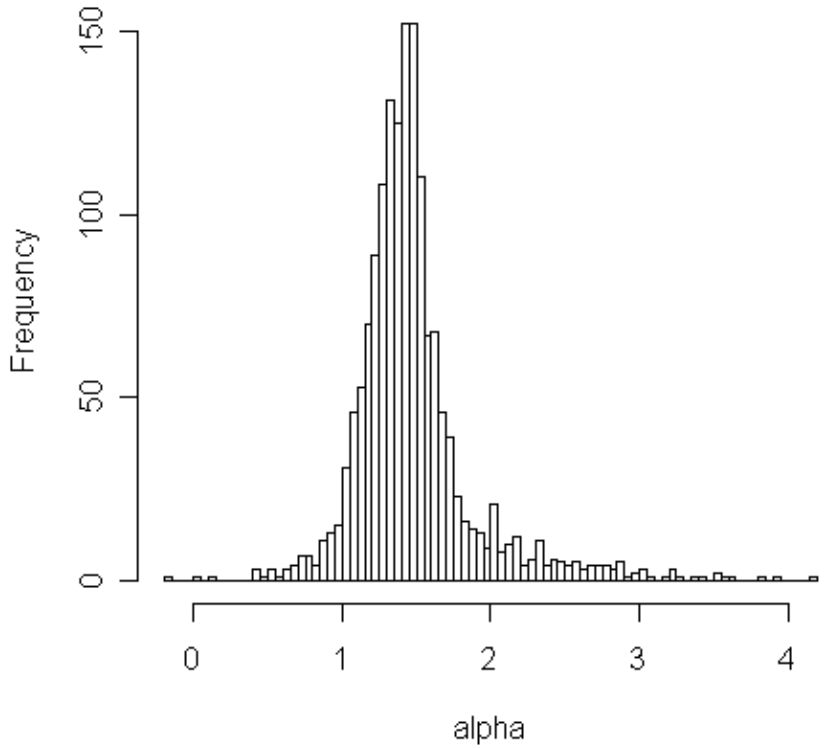


Flickr Tags

- We focus on a set of 1700 tags
- Different growth patterns:
  - bursty (“halloween” or “katrina”)
  - smooth (“landscape” or “bw”)
  - periodic (“moon”)
- For each tag, define an action corresponding to using the tag for the first time.

Social Correlation in Flickr

- Distribution of ® values estimated using maximum likelihood:



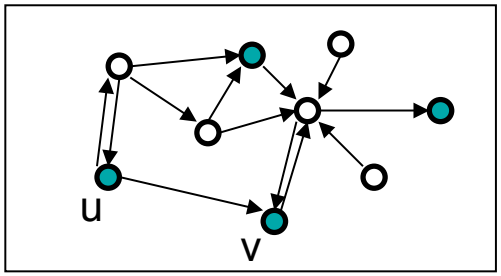
Testing for Influence

- **Simple Idea:** In correlation without influence
  - Who becomes active depends on friends
  - When he becomes active is independent
  - We formalize with a correlation model (omitted)

- Shuffle Test:
  - Compute coefficient ®
  - Shuffle time-stamp of all actions, and re-estimate coefficient ®′
  - If ® ≈ ®′, social influence is ruled out.
  - If ® ≠ ®′, social influence can't be ruled out.

- Edge-Reversal Test:
  - Reverse direction of all edges, and re-estimate ®.

Testing for Influence



Edge-Reversal Test:

- **Simple Idea:**
  - Main idea: assume edge (u → v), where u, v become active
  - If we have influence, u is expected to become active before v
  - If there is no influence, each is equally likely to become active first

- **Test:**
  - Reverse direction of all edges, and re-estimate ®.

Shuffle Test, Theoretical Justification

- Theorem. If the graph is large enough, time-shuffle test rules out the general model of correlation.

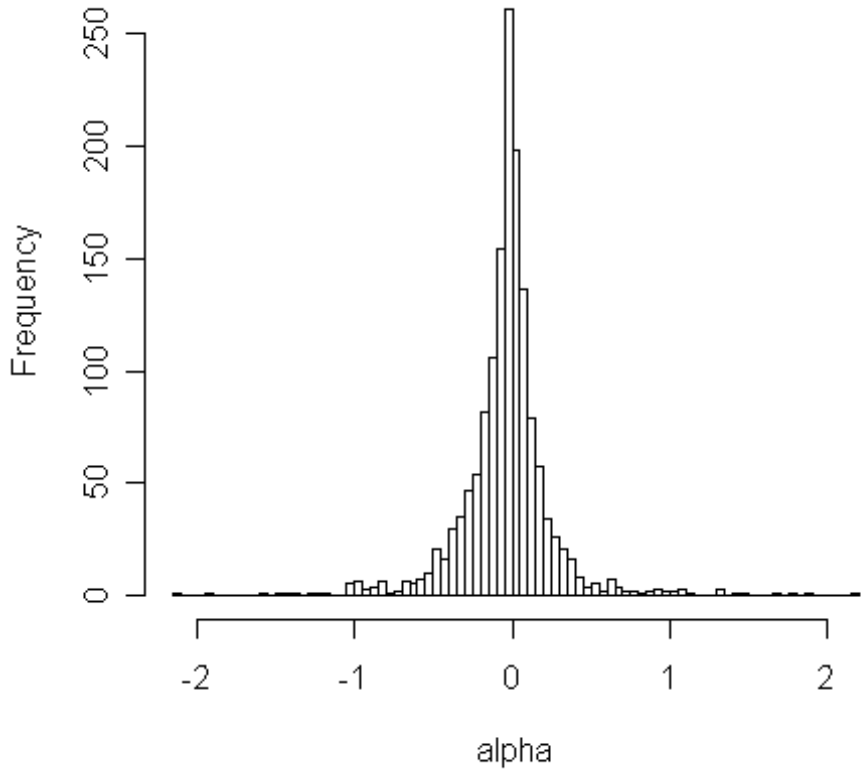
Shuffle Test, Theoretical Justification

- Theorem. If the graph is large enough, time-shuffle test rules out the general model of correlation.
- Intuition: in correlation model, the distribution of the data remains the same if time-stamps are shuffled.
- Challenge: prove concentration.

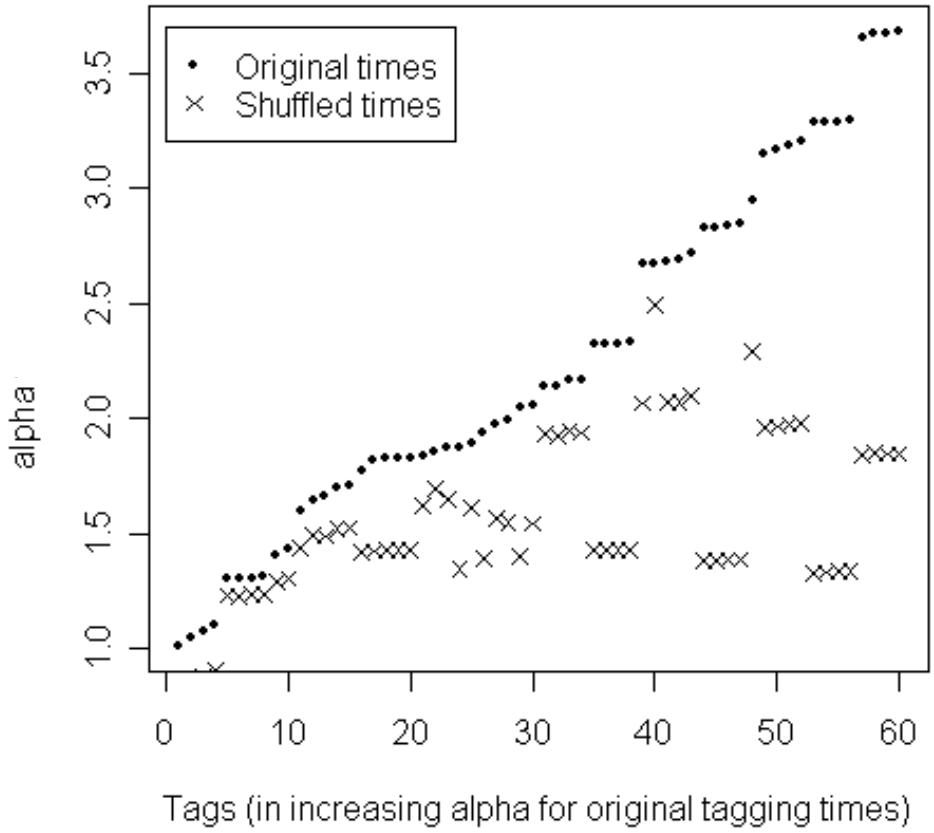
Simulations

- Test methods on randomly generated action data on Flickr network.
- Generate three (network, tag) models:
  - Baseline, no correlation
  - Influence model
  - Correlation model

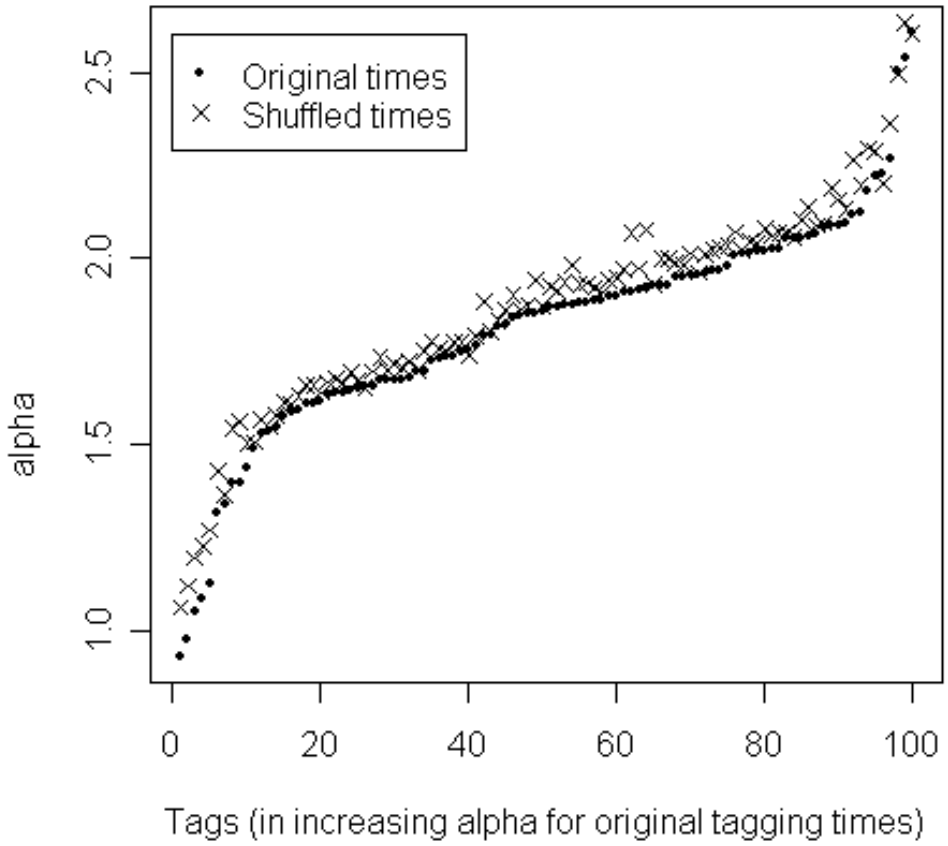
Simulation Results, Baseline



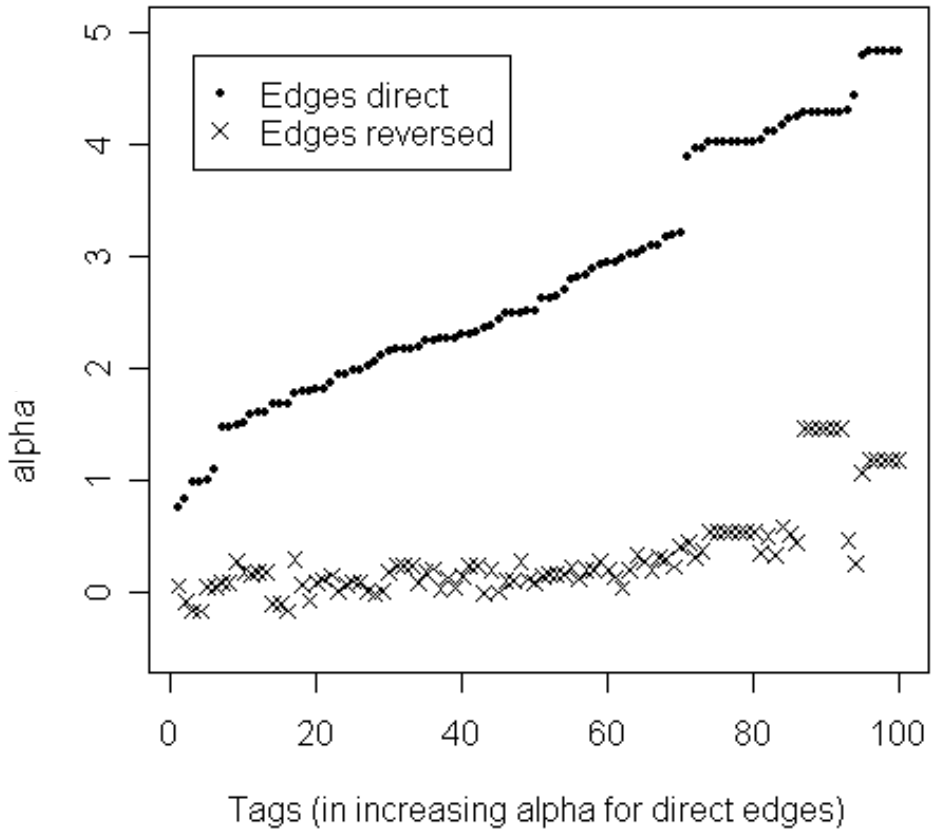
Shuffle Test, Influence Model



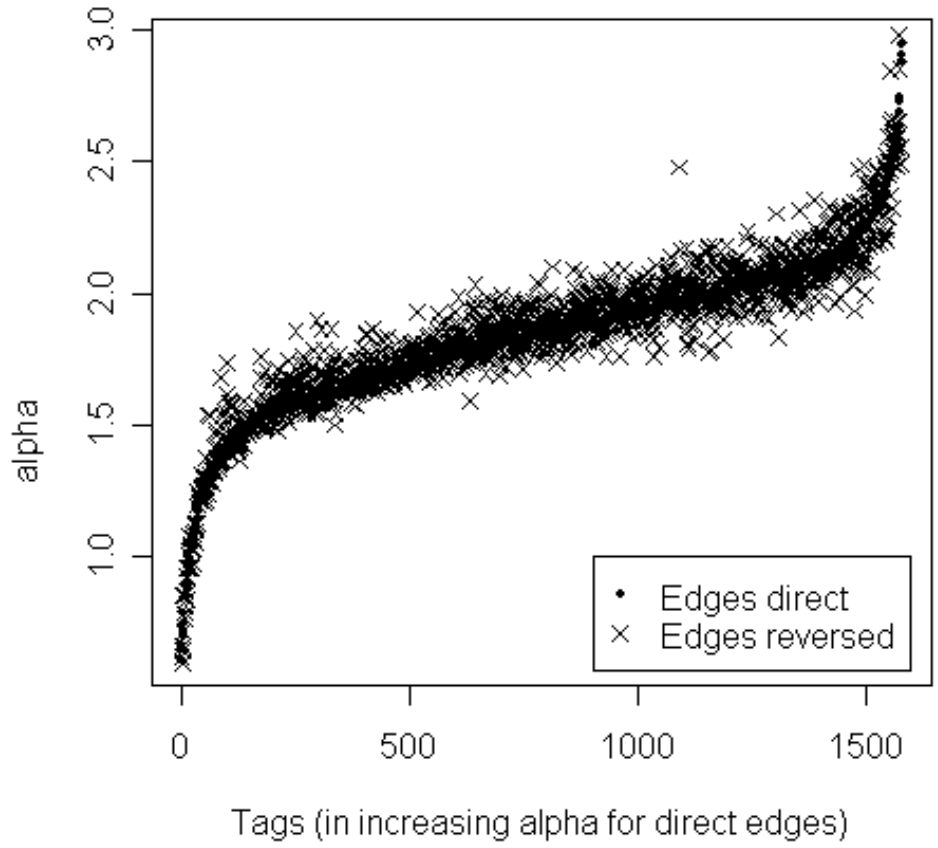
Shuffle Test, Correlation Model



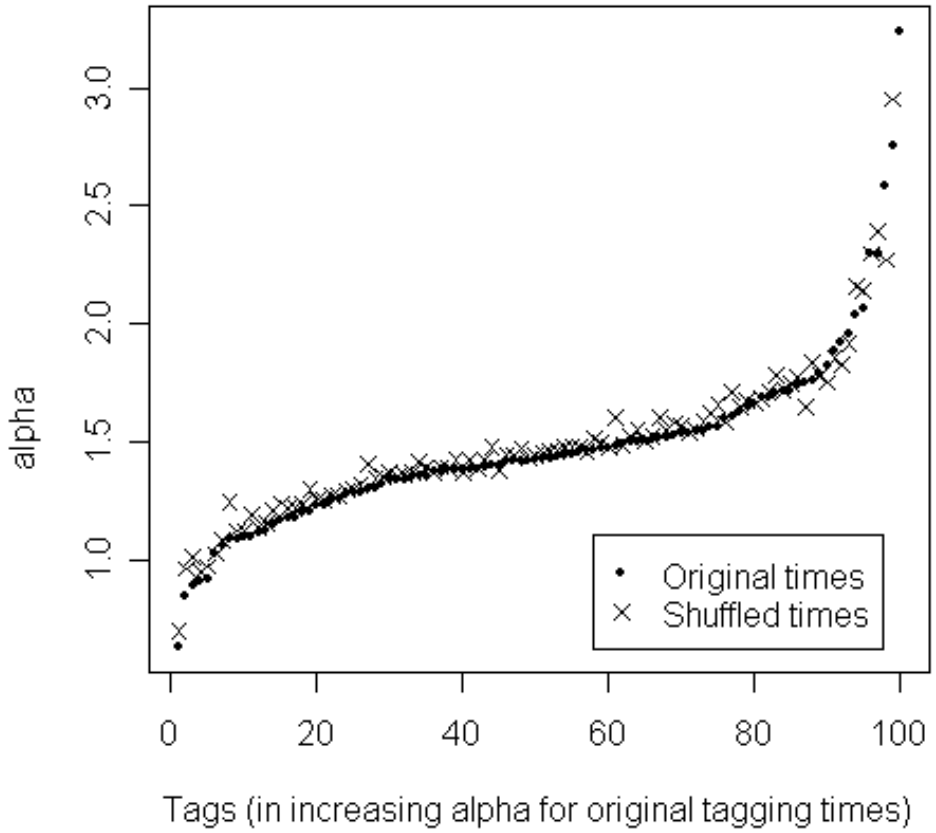
Edge-Reversal Test, Influence Model



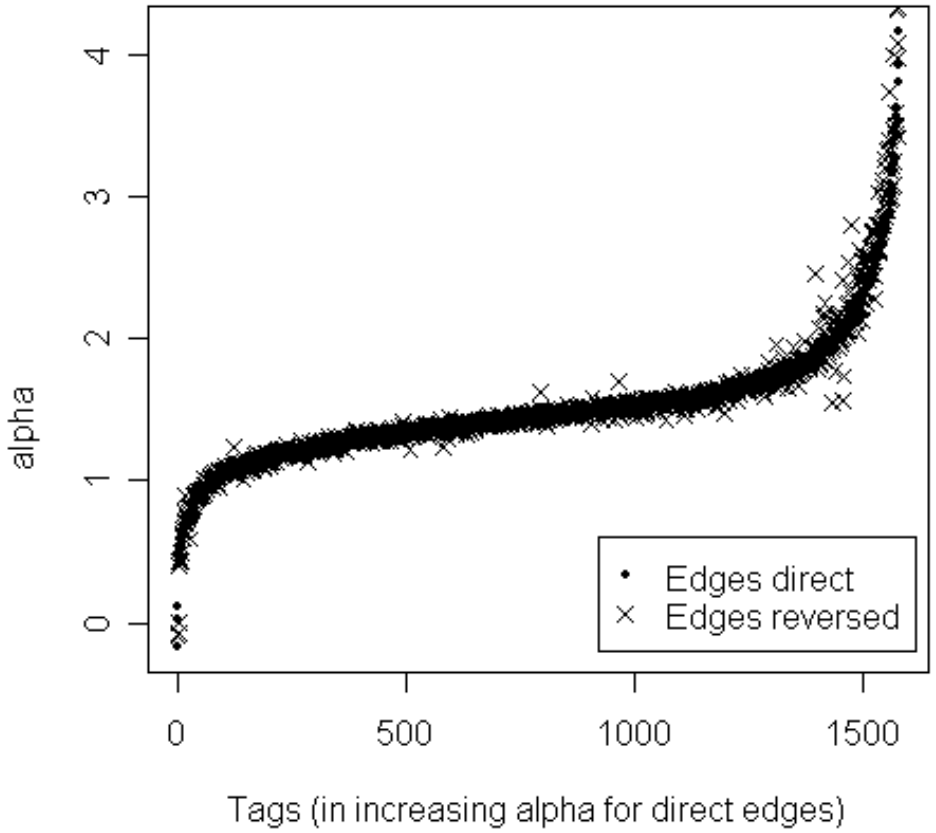
Edge-Reversal Test, Correlation Model



Shuffle Test on Flickr Data



Edge-Reversal Test on Flickr Data



Conclusions

- Summary
  - Defined two models that exhibit correlation, one with and the other without social influence.
  - Two tests to distinguish between them
  - Theoretical justification
  - Simulations suggest that the tests “work” in practice.
  - On Flickr, we conclude that despite considerable correlation, no social influence can be detected.
- Discussion
  - cannot conclusively say there is influence (e.g. flu treatment)
  - still can rule out potential candidates
  - Open: develop more quantitative methods

Thanks!

http://aris.me

aris@cs.brown.edu