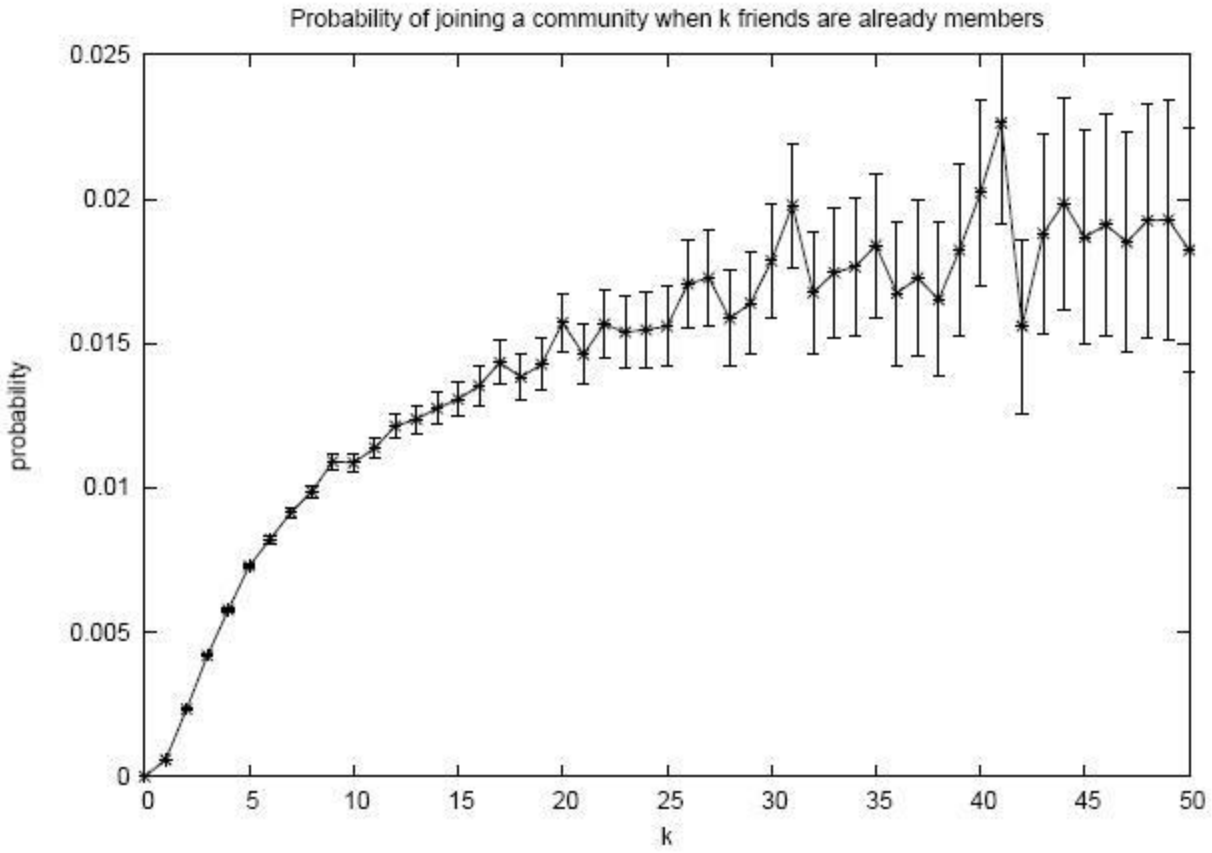
Distinguishing Influence from Correlation in Social Networks

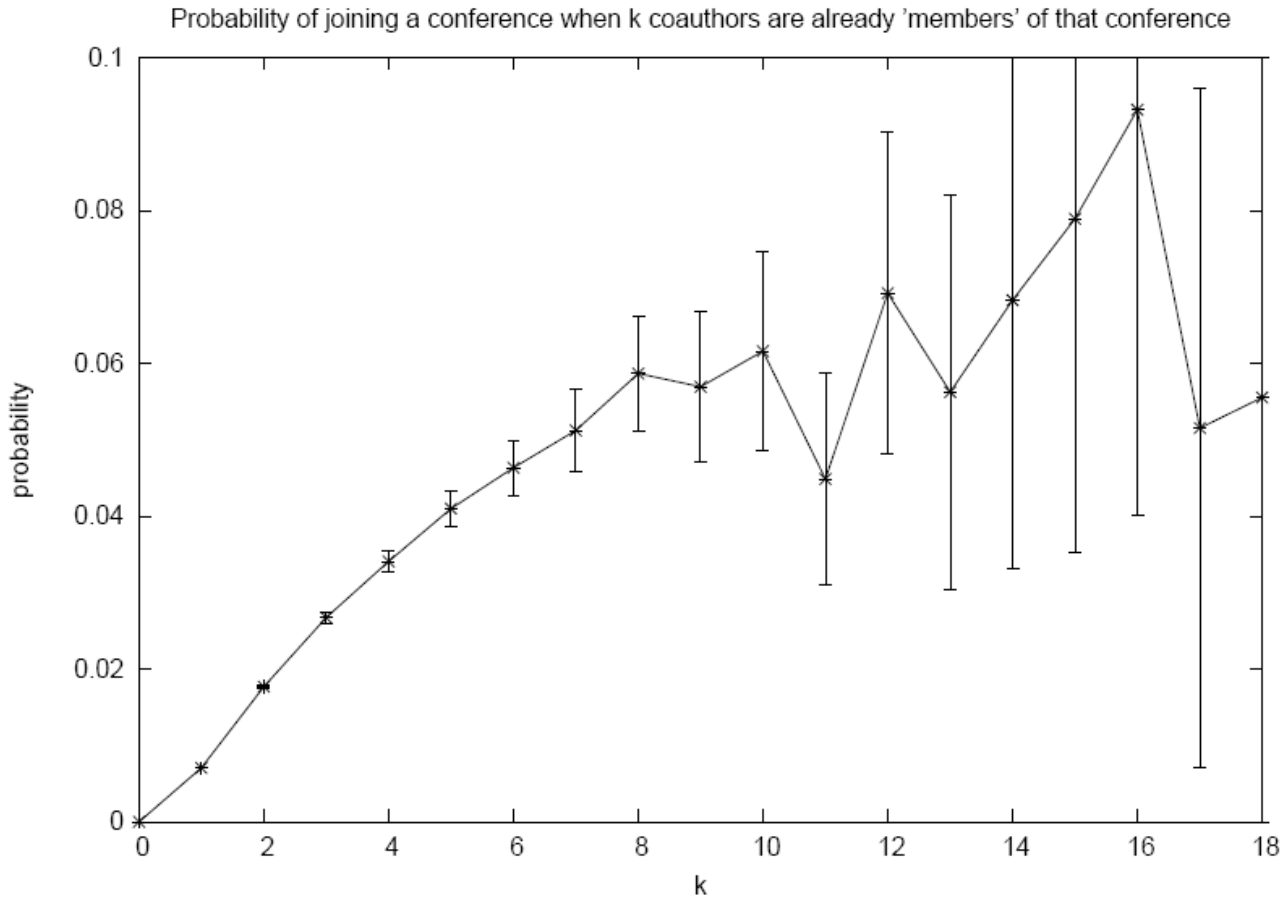
Social correlation

- How similar is the behavior of connected users.
- Previous studies:
  - Joining LiveJournal communities [Backstrom et al.]
  - Publishing in conferences [Backstrom et al.]
  - Tagging vocabulary on flickr [Marlow et al.]
  - Adoption of paid VoIP service in IM
  - Offline: Smoking habits of teenagers
  - …

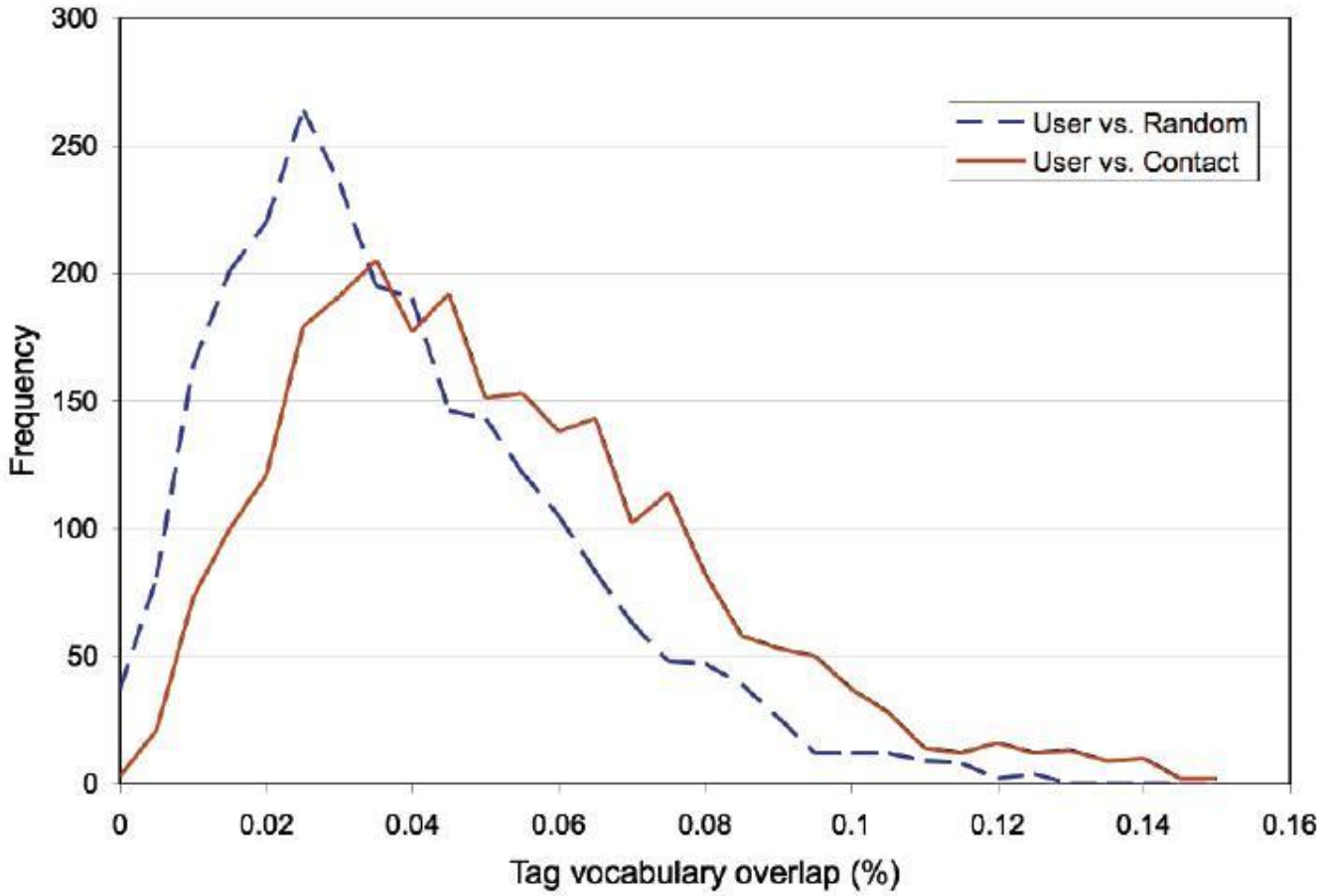
Joining communities [Backstrom et al]



Publishing in conferences



Flickr tag vocabulary [Marlow et al.]



Sources of correlation

- Social influence: 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: 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: friends are more likely to live in the same area, thus attend and take pictures of similar events, and tag them with similar tags

Social influence

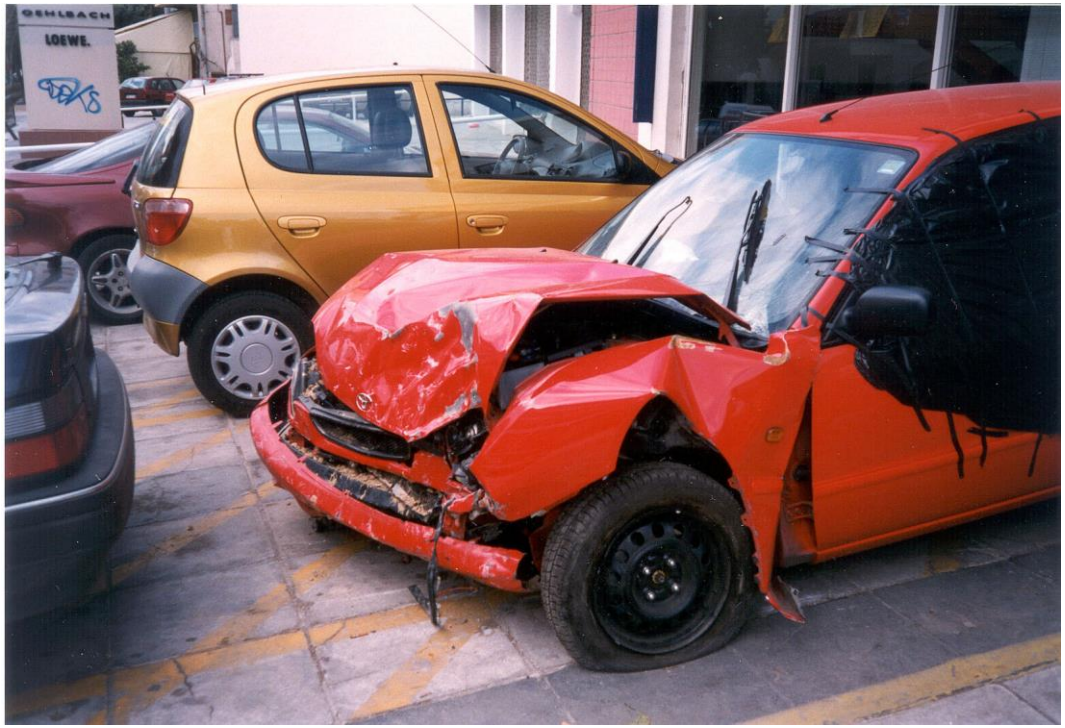
- Focus on a particular "action" A.
  - E.g.: buying a product, joining a community, publishing in a conference, using a particular tag, using the VoIP service, …
- An agent who performs A is called "active"
- x has influence over y if x performing A increases the likelihood that y performs A.

- Distinguishing factor: causality relationship

Causation vs. Correlation

- What we try to do is essentially distinguish causation from correlation.
- Common mistake, especially by journalists:
  - People who drink more coffee live longer
  - People who drive red cars create more accidents

Causation vs. Correlation

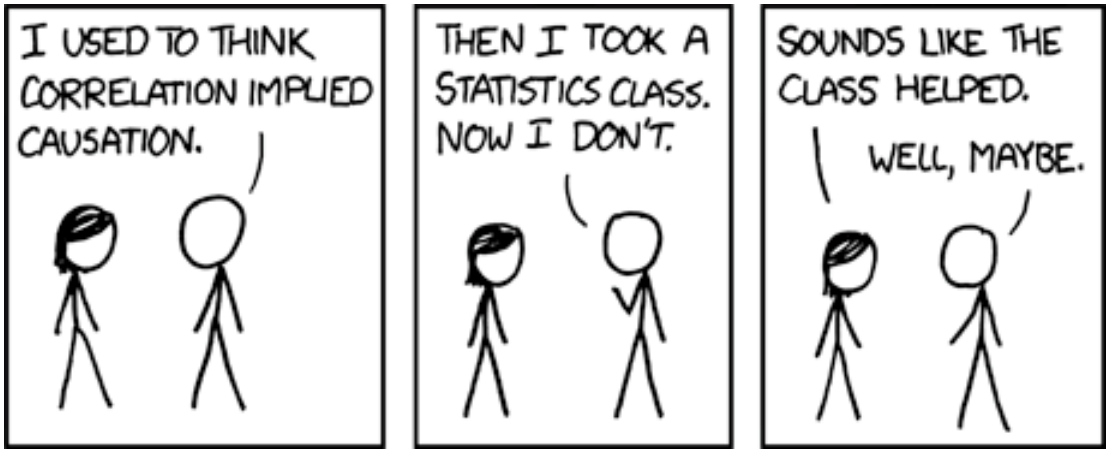


Causation vs. Correlation

- What we try to do is essentially distinguish causation from correlation.
- Common mistake, especially by journalists:
  - People who drink more coffee live longer
  - People who drive red cars create more accidents
  - Eating pizza "cuts cancer risk“
  - People who go to school, live longer

Causation vs. Correlation

- What we try to do is essentially distinguish causation from correlation.
- Common mistake, especially by journalists:
  - People who drink more coffee live longer
  - People who drive red cars create more accidents
  - Eating pizza "cuts cancer risk“
  - People who go to school, live longer



Causation vs. Correlation

The Washington Times
America's Newspaper

News Policy Commentary Sports Sponsored Events Video/Podcasts Games

TRENDING: GAZA | ISRAEL | JOE BIDEN | DONALD TRUMP | HAMAS | UKRAINE | ADAM SCHIFF | BENJAMIN NETANYAHU | KATE | LISBON

HOME \ NEWS \ NATIONAL

Pizza reduces cancer risk: study

Print

By Cheryl K. Chumley - *The Washington Times - Tuesday, April 30, 2013*

Eat more pizza. That's the message from Italian researchers who say eating more of the doughy pie can actually cut the chance for certain cancers.

The secret?

It's likely in the tomatoes, researches said, according to a BBC report.

Follow Us

Search

SEE MORE VIDEOS

Causation vs. Correlation

Int. J. Cancer: **107,** 283–284 (2003)

Publication of the International Union Against Cancer

DOES PIZZA PROTECT AGAINST CANCER?

Silvano GALLUS[1]*, Cristina BOSETTI[1], Eva NEGRI[1], Renato TALAMINI[2], Maurizio MONTELLA[3], Ettore CONTI[4], Silvia FRANCESCHI[5] and Carlo LA VECCHIA[1,6]

[1]*Istituto di Ricerche Farmacologiche "Mario Negri," Milan, Italy*
[2]*Centro di Riferimento Oncologico, Aviano (PN), Italy*
[3]*Istituto Tumori "Fondazione Pascale," Cappella dei Cangiani, Naples, Italy*
[4]*Istituto "Regina Elena" per lo Studio e la Cura dei Tumori, Rome, Italy*
[5]*International Agency for Research on Cancer, Lyon, France*
[6]*Istituto di Statistica Medica e Biometria, Università degli Studi di Milano, Milan, Italy*

We analyzed the potential role of pizza on cancer risk, using data from an integrated network of case-control studies conducted in Italy between 1991 and 2000. Cancer sites were: oral cavity and pharynx (598 cases), esophagus (304 cases), larynx (460 cases), colon (1,225 cases) and rectum (728 cases). Controls were 4,999 patients admitted for acute, non-neoplastic conditions to the same hospital network as cases. Odds ratios for regular pizza consumers were 0.66 (95% confidence interval, CI = 0.47–0.93) for oral and pharyngeal cancer, 0.41 (95% CI = 0.25–0.69) for oesophageal, 0.82 (95% CI = 0.56–1.19) for laryngeal, 0.74 (95% CI = 0.61–0.89) for colon and 0.93 (95% CI = 0.75–1.17) for rectal cancer. Pizza appears therefore to be a favorable indicator of risk for digestive tract neoplasms in this population.

Key words: *digestive tract cancers; lycopene; pizza; risk factors*

Pizza is one of the best known and most widespread Italian foods, and it is said to be the most common generic commercial sign of Italy worldwide. Investigating and quantifying any potential role of pizza on cancer risk seems to be a curious issue, but may well have interesting implications in respect to dietary advice in Italy as well as elsewhere.

Limited and inconclusive information is available on the potential influence of pizza, however, as a food item or as an indicator of any specific dietary pattern, on cancer risk. An inverse trend in

jects' usual diet before diagnosis (or hospital admission) was investigated using a validated 78-item food frequency questionnaire[8,9] that included a specific question on pizza. For the present analyses, pizza eating was classified in 3 categories: non eaters (<1 portion of pizza/month), occasional eaters (1–3 portions/month) and regular eaters (1 portion of pizza or more/week).

OR and the corresponding 95% CI, for subsequent levels of pizza eating were derived by unconditional multiple logistic regression models, including terms for age, gender, study center, education, alcohol and tobacco consumption, energy intake, body mass index and for colon and rectum, a measure of physical activity.

RESULTS

Table I shows the distribution of cases and controls according to pizza consumption and the corresponding multivariate ORs. Compared to non-pizza-consumers, the multivariate ORs for pizza eaters (≥1 portion/month) were 0.73 for oral cavity and pharynx, 0.53 for esophagus, 0.85 for larynx, 0.81 for colon and 0.88 for rectum. Corresponding ORs for regular pizza eaters (≥1 portion/week) were 0.66 for oral and pharyngeal, 0.41 for oesophageal, 0.82 for laryngeal, 0.74 for colon and 0.93 for rectal cancer. The trends in risk were significant for oral and pharyngeal, esophageal and colon cancers.

Causation vs. Correlation

The New York Times

THE NEW AGE

A Surprising Secret to a Long Life: Stay in School

Irma Lara, 75, who came to the United States from Mexico when she was 26, spends time exercising at a community center in Texas.
Michael Stravato for The New York Times

By Gina Kolata
Jan. 3, 2007

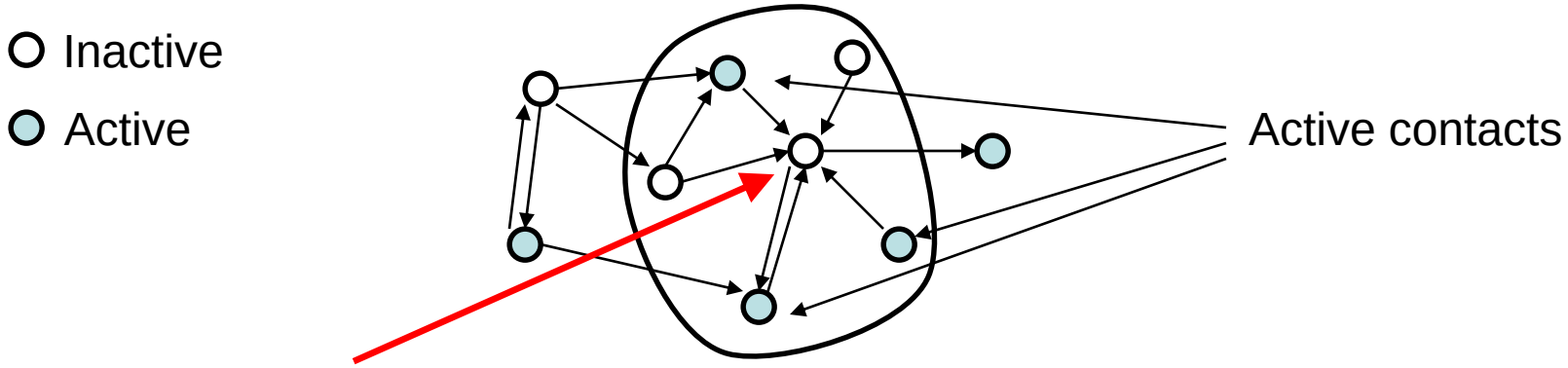
James Smith, a health economist at the RAND Corporation, has heard a variety of hypotheses about what it takes to live a long life — money, lack of stress, a loving family, lots of friends. But he has been a skeptic.

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: for designing a system to induce a particular behavior, e.g.:
  - vaccination strategies (random, targeting a demographic group, random acquaintances, etc.)
  - viral marketing campaigns

Influence Model

- Graph (static or dynamic)
- Edge (u,v): Node u can influence node v
- Discrete time: t = 0, 1, 2, …, T
- 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 as the explanatory variable. I.e.,

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

- Coefficient x measures social correlation.

Model – Influence probability

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

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

with x as the explanatory variable. I.e.,

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

- Coefficient x measures social correlation.

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 as the explanatory variable. I.e.,

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

- Coefficient x measures social correlation.

Measuring social correlation

- Given data, we compute the maximum likelihood estimate for parameters α and β.

- Let Y_x = # pairs (user u, time t) where u is not active and has x active friends at the beginning of time step t, and becomes active in this step.
- Let N_x = …… does not become active in this step.

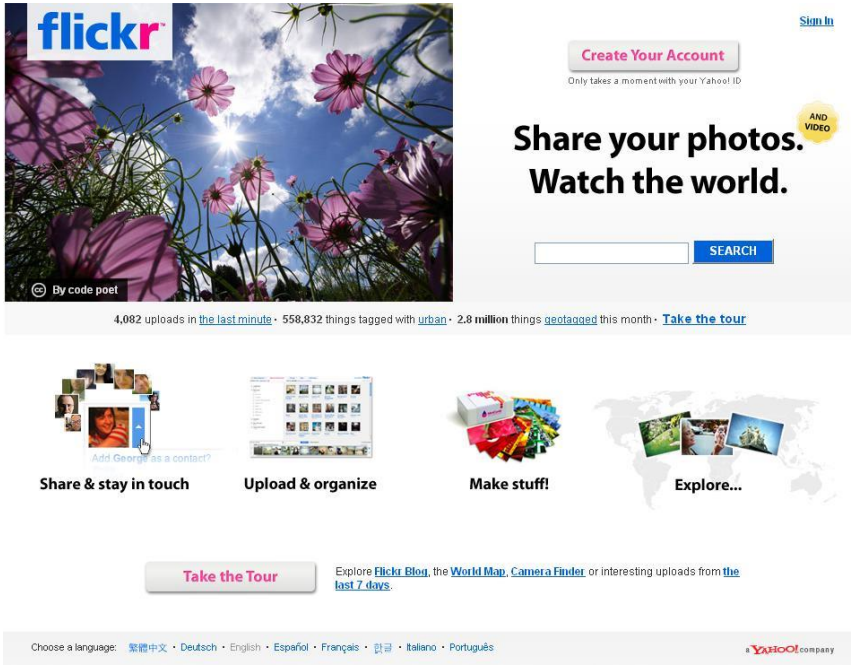
- Find α, β to maximize the likelihood function:

$$f(\alpha, \beta, Y_x, N_x) = \prod_x p(x)^{Y_x} \big(1 - p(x)\big)^{N_x}$$

- For convenience, we cap x at a value R.

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”.
  - 2.8M directed edges, 28.5% of edges not mutual.
  - Size of giant component ~160K



mmahdian's photostream pro

portrait

Uploaded on Apr 7, 2008

2 notes / 7 comments

graffiti

"None are more hoplessly enslaved than those who falsely believe they are free." graffiti...

Uploaded on Feb 20, 2008

4 comments

golden gate

this photo was taken by mistake! i took the photo after changing lens, and the lens was...

roja

Uploaded on Dec 3, 2007

2 comments

iran

19 photos

flowers

12 photos

funny pix

4 photos

faves

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

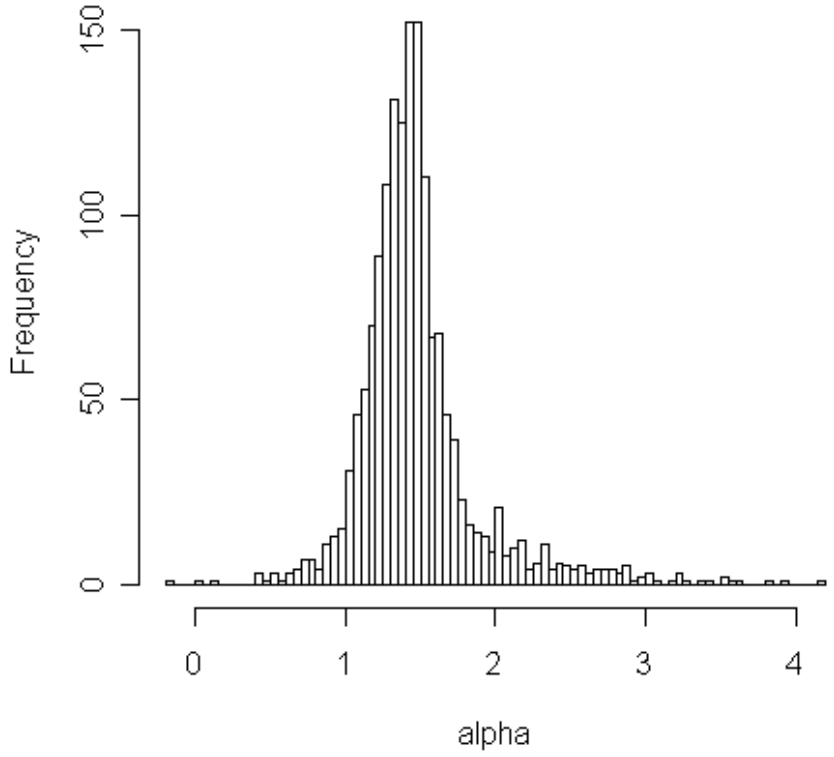
Additional Information

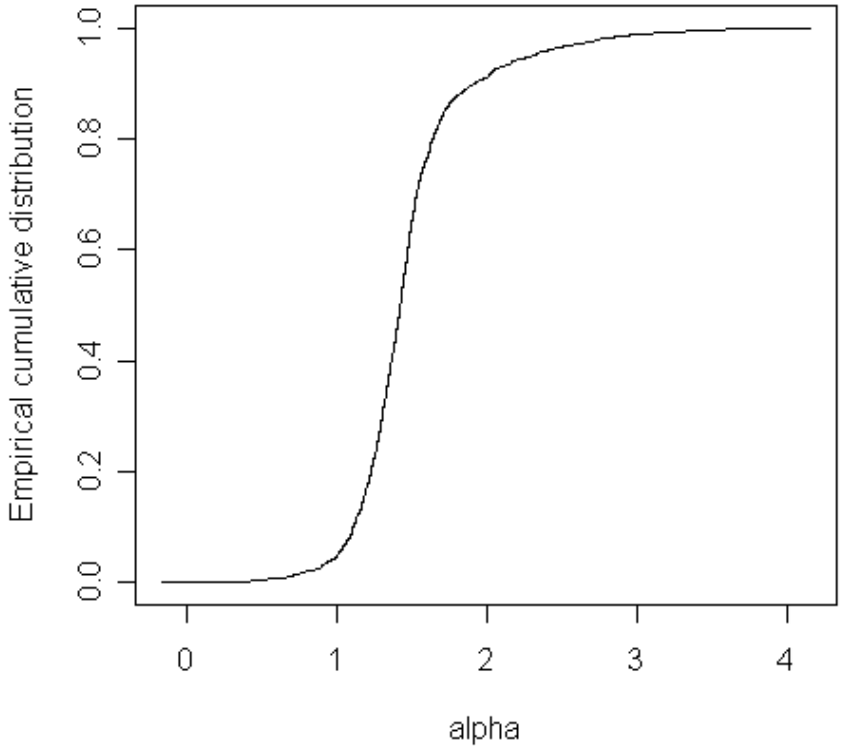
Flickr tags

- ~10K tags
- We focus on a set of 1700
- 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.

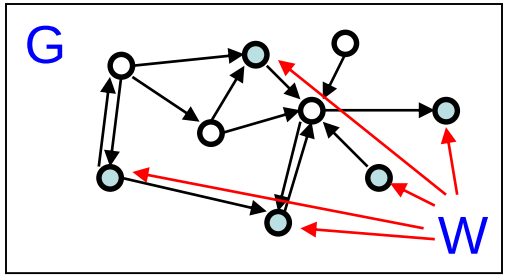
Social correlation in flickr

- Distribution of alpha (α) values estimated using maximum likelihood:



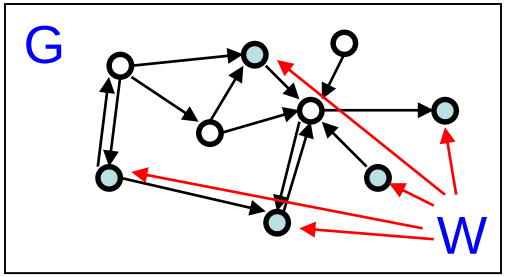


Distinguishing influence



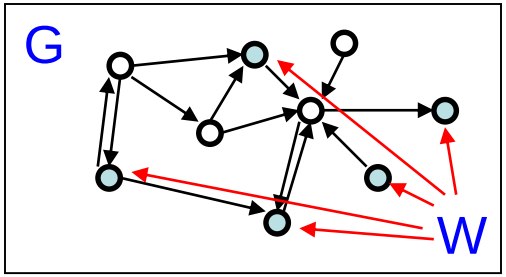
- Recall: graph G, set W of active nodes
- Influence model
  - First G is selected
  - Then W is picked from a distribution depending on G

Distinguishing influence



- Noninfluence models
  - Homophily (Similar individuals are more likely to become friends):
    - First W is picked, then G is picked from a distribution that depends on W
  - Confounding factors (External influence from elements in the environment):
    - Both G and W are picked from distributions that depend on another var X

Distinguishing influence



- Generally, we consider this correlation model:
  - (G,W) are selected from a joint distribution
  - Each agent in W picks an activation time i.i.d. from a distribution on [0,T]

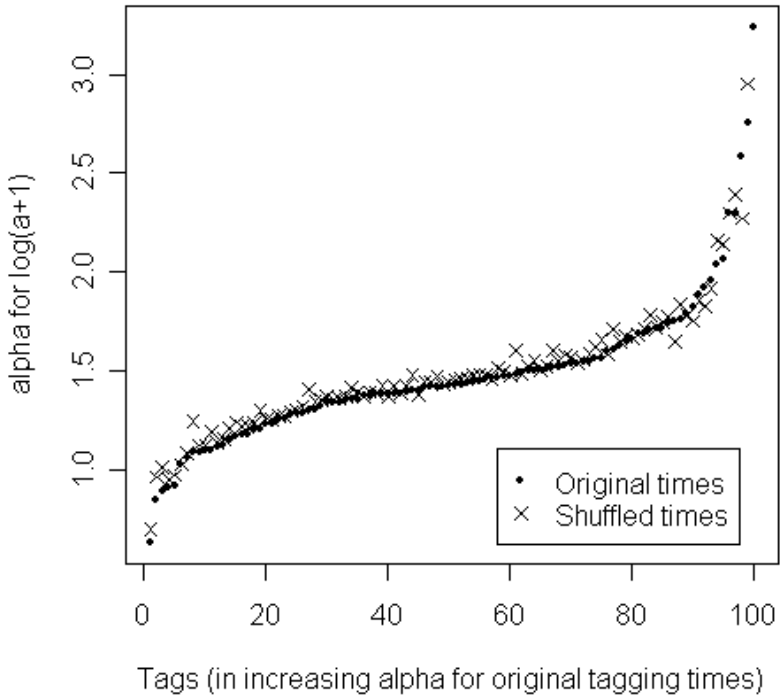
Testing for influence

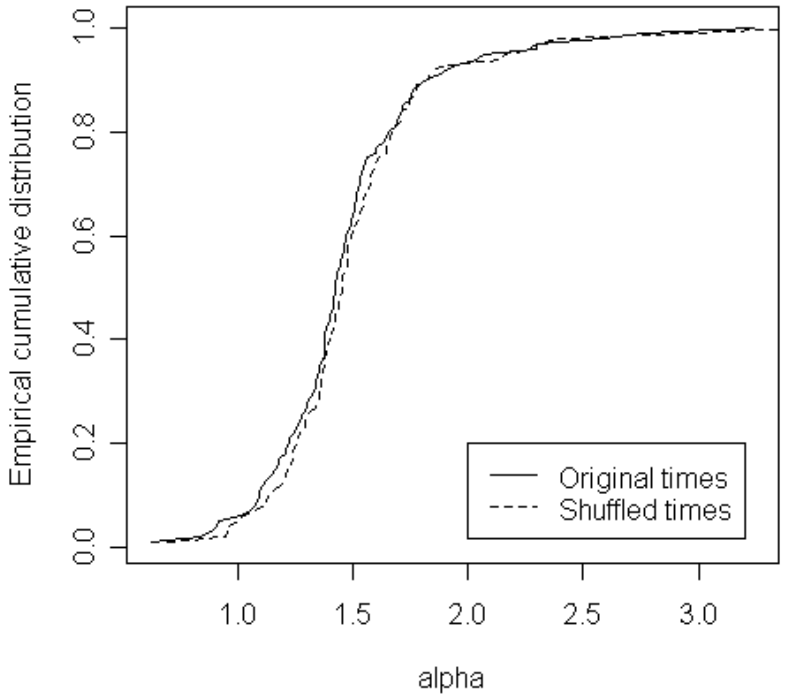
- **Simple idea:** even though an agent's probability of activation can depend on friends, her timing of activation is independent
- Shuffle Test: re-shuffle the time-stamp of all actions, and re-estimate the coefficient α. If different from original α, social influence can't be ruled out.

Testing for influence

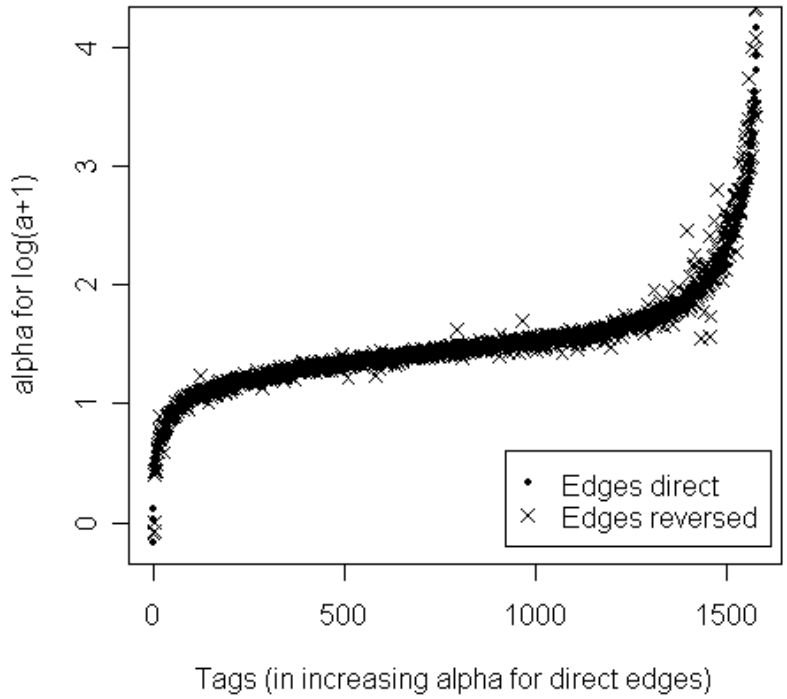
- **Simple idea:** even though an agent's probability of activation can depend on friends, her timing of activation is independent
- Shuffle Test: re-shuffle the time-stamp of all actions, and re-estimate the coefficient α. If different from original α, social influence can't be ruled out.
- Edge-Reversal Test: reverse the direction of all edges, and re-estimate α.

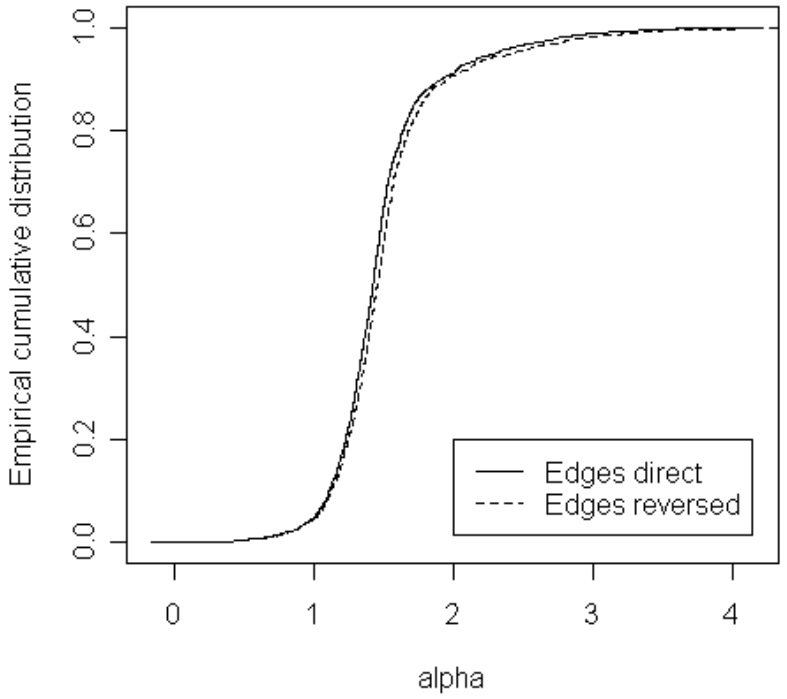
Shuffle test on Flickr data





Edge-reversal test on Flickr data

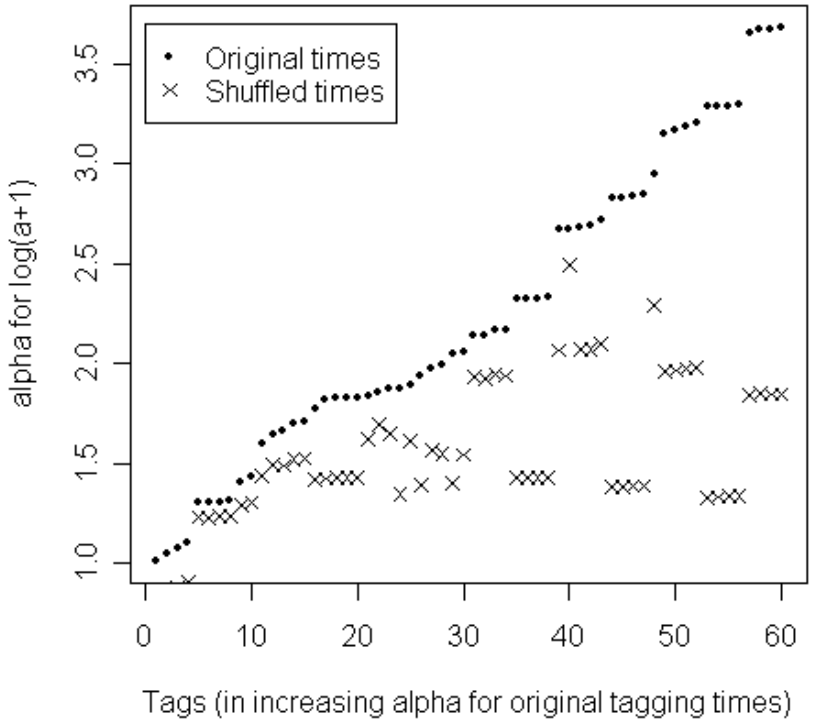


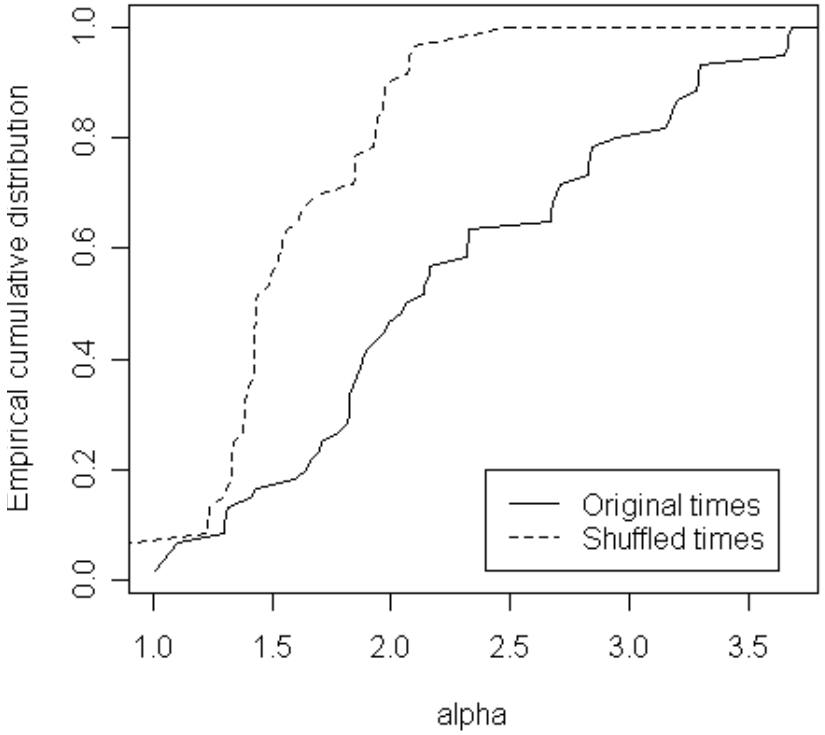


Simulations

- Run the tests on randomly generated action data on flickr network.
- Baseline: no-correlation model, actions generated randomly to follow the pattern of one of the real tags, but ignoring network
- Influence model: same as described, with a variety of (α,β) values
- Correlation model: pick a # of random centers, let W be the union of balls of radius 2 around these centers.

Shuffle test, influence model





Edge-reversal test, influence model