# Probabilistic models for Social Influence Propagation

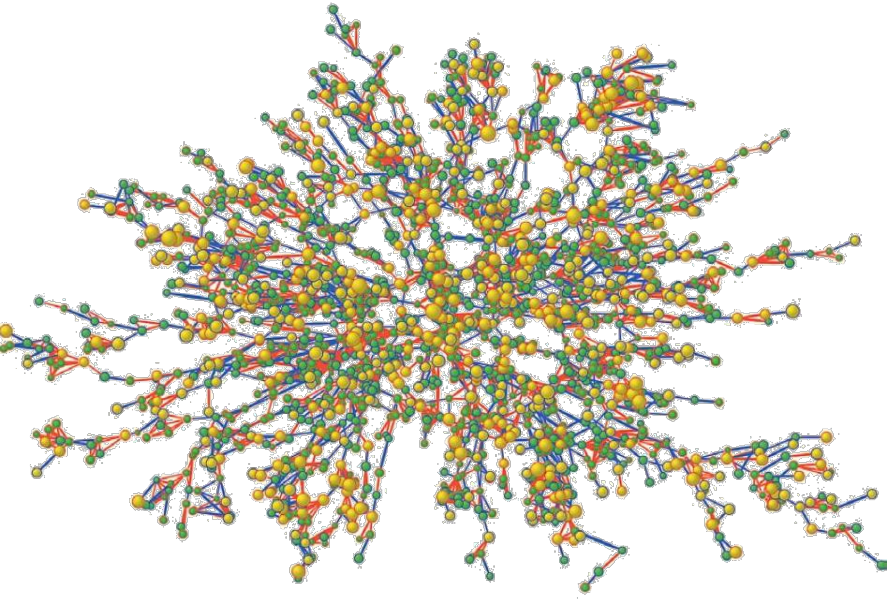
Giuseppe Manco

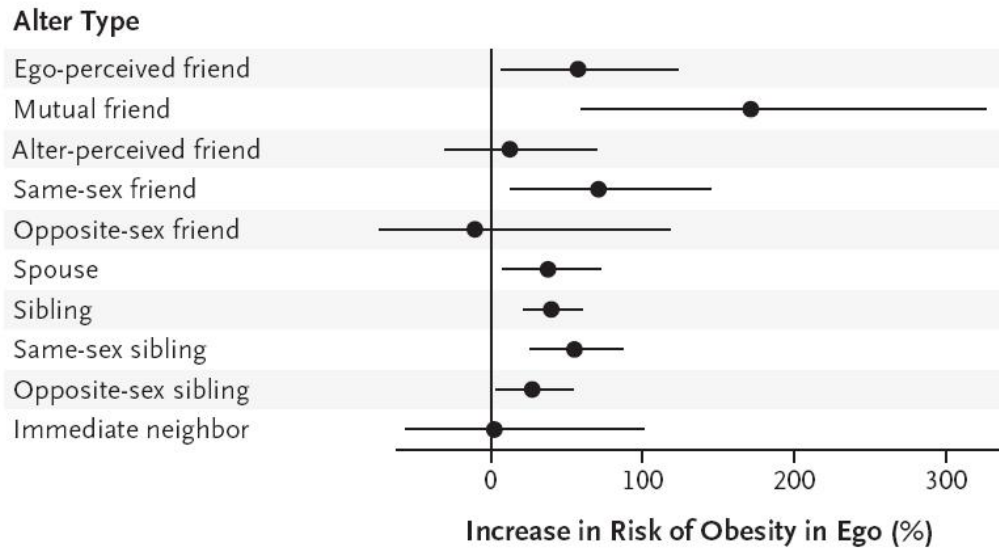
ICAR-CNR Rende, Italy
manco@icar.cnr.it

# The Spread of Obesity in a Large Social Network over 32 Years

Christakis and Fowler, New England Journal of Medicine, 2007

Data set: 12,067 people from 1971 to 2003, 50K links






Obese Friend → 57% increase in chances of obesity

Obese Sibling → 40% increase in chances of obesity

Obese Spouse → 37% increase in chances of obesity

# Influence or Homophily?

(we don't care in this paper)

## Homophily

tendency to stay together with people similar to you

*"Birds of a feather flock together"*

---

## Social influence

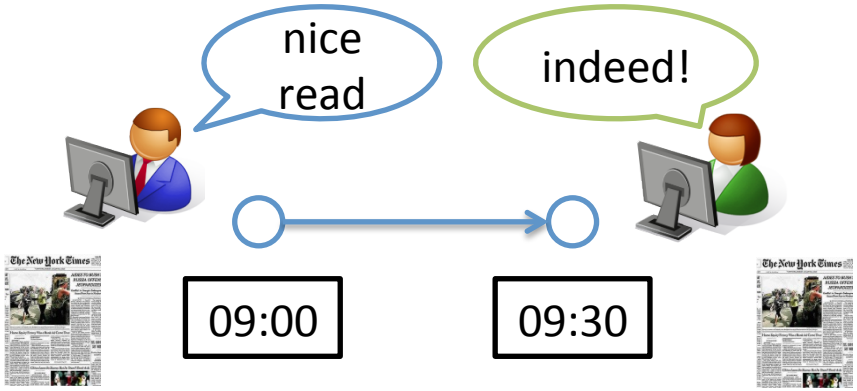
a force that person A (i.e., the influencer) exerts on person B
to introduce a change of the behavior and/or opinion of B

Influence is a causal process

Problem: How to distinguish social influence from homophily and other factors of correlation

Crandall et al. (KDD'08) *"Feedback Effects between Similarity and Social Influence in Online Communities"*
Anagnostopoulos et al. (KDD'08) *"Influence and correlation in social networks"*
Aral et al. (PNAS'09) *"Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks"*
Myers et al. (KDD'12) *"Information Diffusion and External Influence in Networks"*

# Influence in on-line social networks




users perform actions

post messages, pictures, video

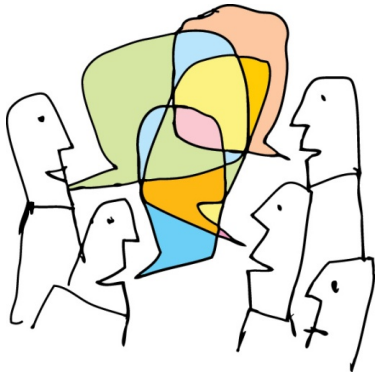
buy, comment, link, rate, share, like, retweet

users are connected with other users

interact, influence each other

actions propagate

# Social Influence Marketing
# Viral Marketing
# WOMM



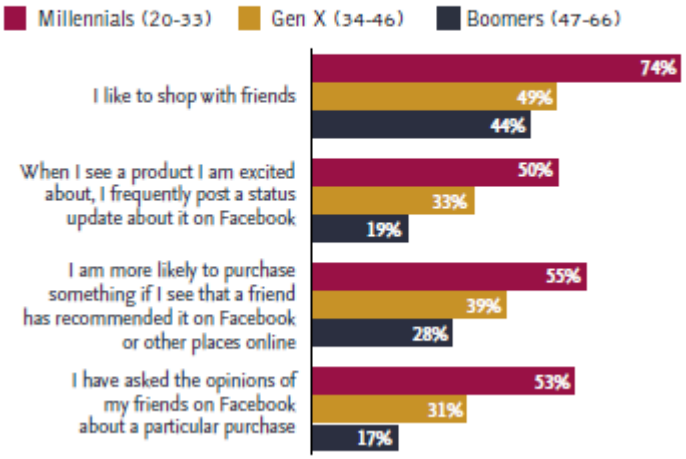
IDEA: exploit social influence for marketing

Basic assumption: word-of-mouth effect, thanks to which actions, opinions, buying behaviors, innovations and so on, propagate in a social network.

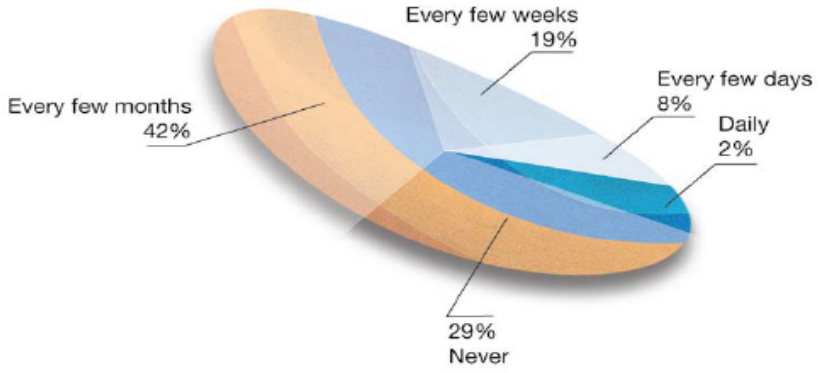
Target users who are likely to produce word-of-mouth diffusion, thus leading to additional reach, clicks, conversions, or brand awareness

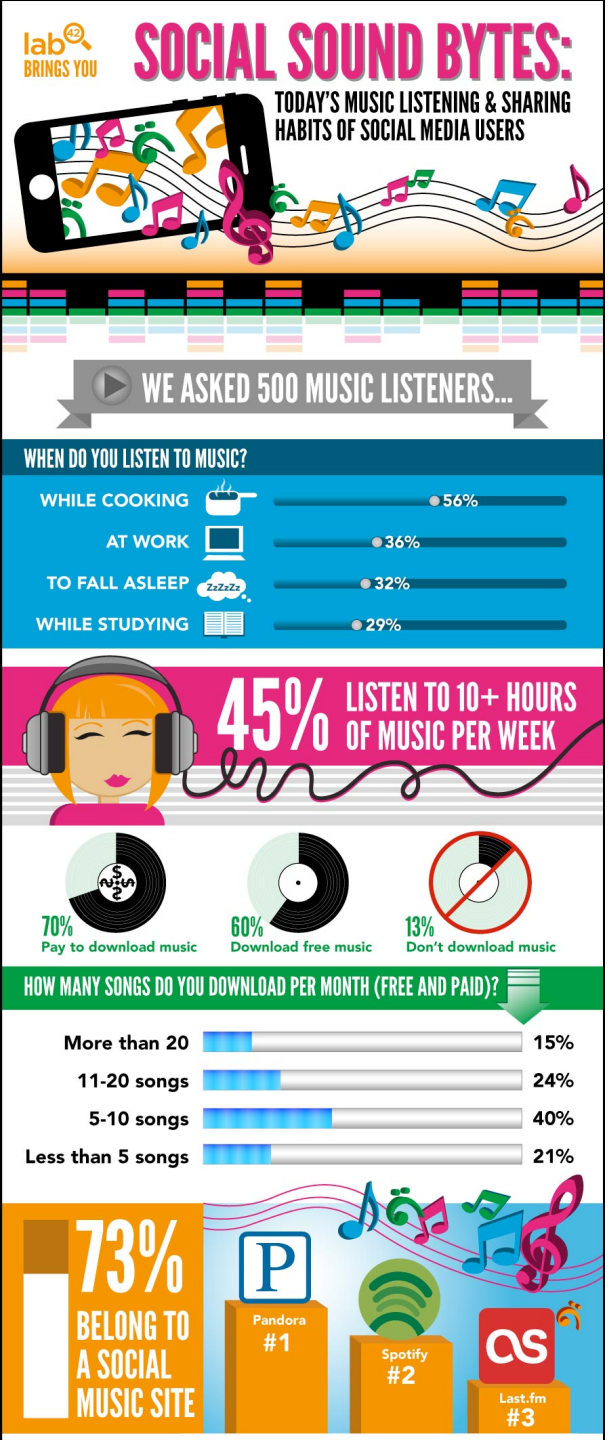
Target the influencers

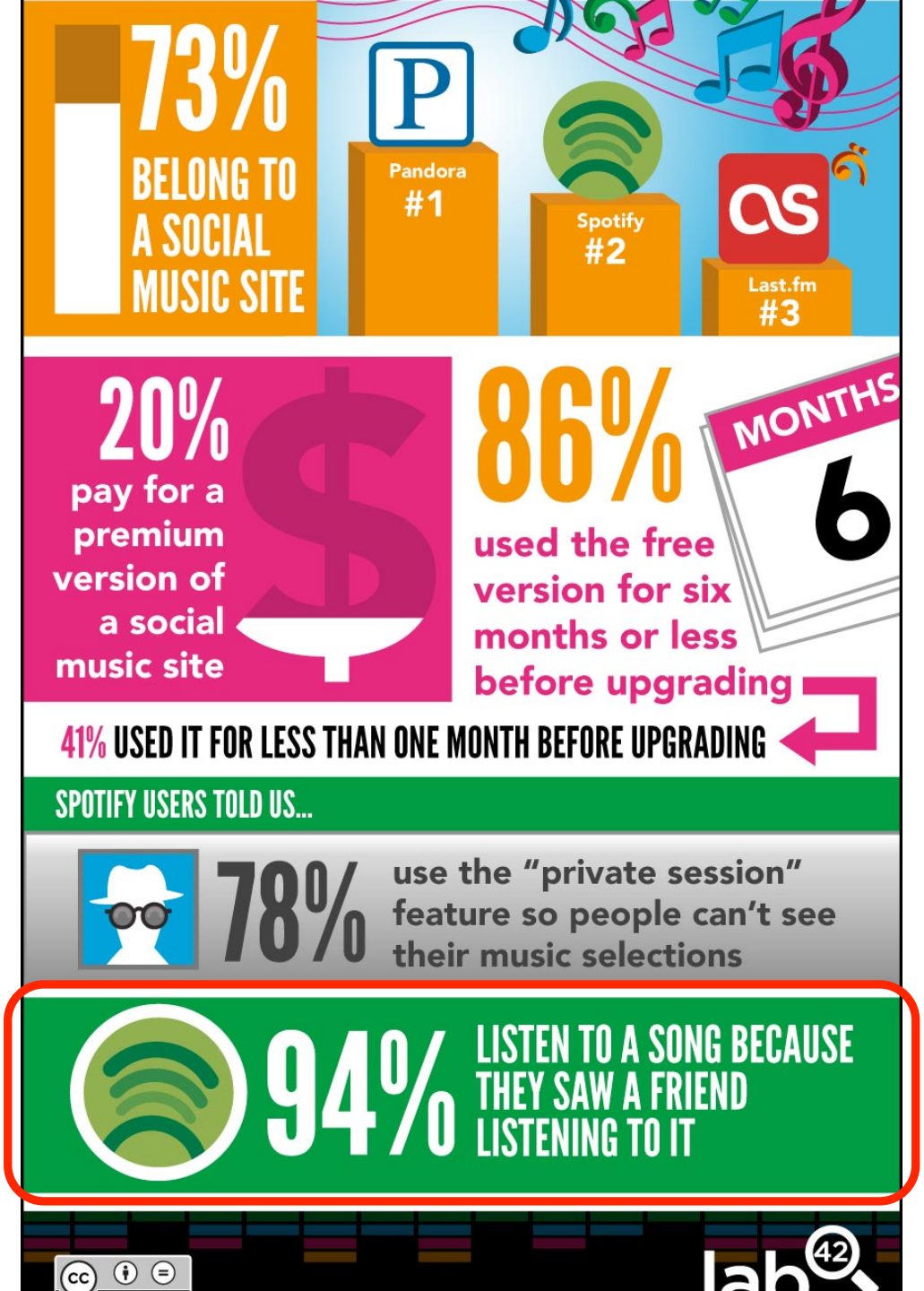
Sharing and social influence




How frequently do you share recommendations online?





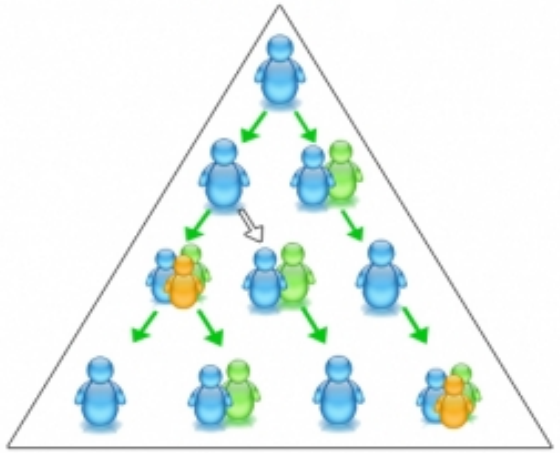




http://mashable.com/2012/08/12/social-music-listening/

# Viral Marketing and Influence Maximization

Business goal (Viral Marketing): exploit the "word-of-mouth" effect in a social network to achieve marketing objectives through self-replicating viral processes

Mining problem: find a seed-set of influential people such that by targeting them we maximize the spread of viral propagations



Hot topic in Data Mining research since 10 years:

Domingos and Richardson *"Mining the network value of customers"* (KDD'01)

Domingos and Richardson *"Mining knowledge-sharing sites for viral marketing"* (KDD'02)

Kempe et al. *"Maximizing the spread of influence through a social network"* (KDD'03)

Lady Gaga @ladygaga

The Economist @TheEconomist

Benedict XVI @Pontifex

CNN @cnn

Bill Gates @billgates

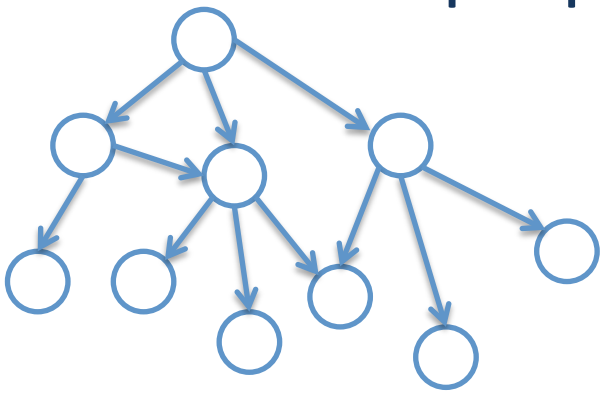
# Who are the most influential users in a SN?

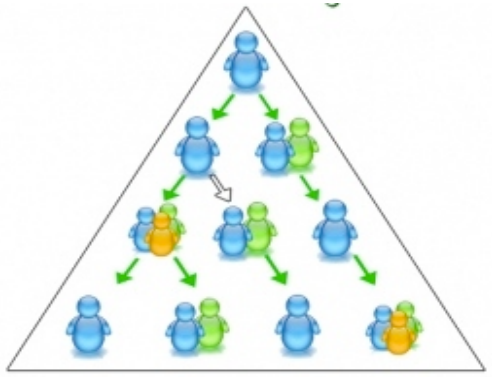
TomCruise.com @tomcruise

Barack Obama @barackobama

Leonardo DiCaprio @LeoDiCaprio

Who are the people to target such that by targeting them we maximize the spread of viral propagations?









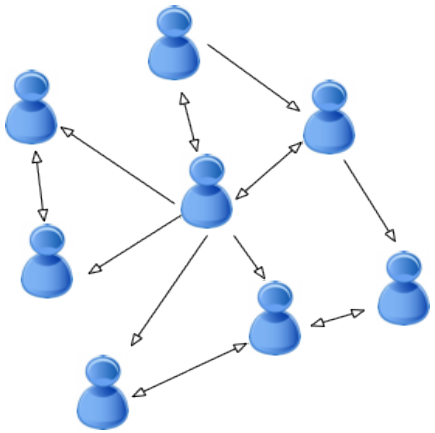
# Users authoritativeness, expertise, trust and influence are topic-dependent!!!




# Topic-aware Social Influence Propagation Models

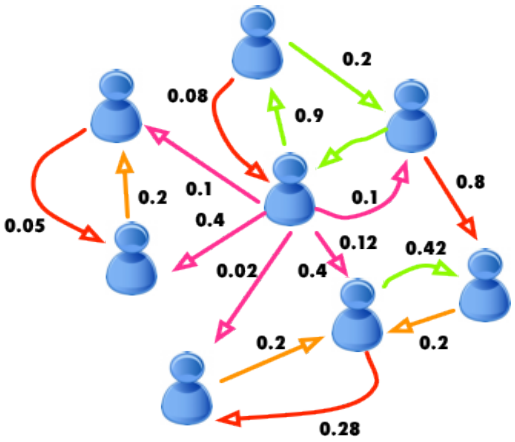
## Input:

Social Network + Propagation Log
+ number of topics K



## Output:

Topic-Aware Influence Strength




## Applications:

**Better users' behavior modeling**

**Detection of influent authorities for different topics**

**Design of targeted viral marketing strategies**

# Topic-aware influence models

Barbieri, Bonchi, Manco (ICDM'12)

# Simple Topic-Aware Influence Propagation Models:

**For each item that propagates in the network, we have a distribution over topics**

$$\gamma_i^z = P(Z = z|i)$$

**User-to-user influence probabilities depend on the topic**

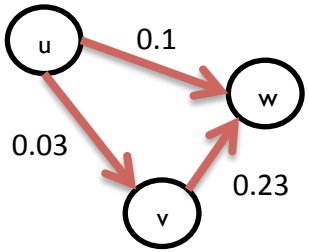
**Topic-Aware Independent Cascade (TIC)**

$$p_{v,u}^i = \sum_{z=1}^{K} \gamma_i^z p_{v,u}^z$$

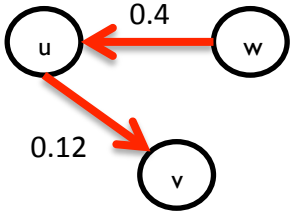
**Topic-Aware Linear Threshold model (TLT)**

$$W_i^t(u) = \sum_{z=1}^{K} \sum_{v \in \mathcal{F}_i(u,t)} \gamma_i^z p_{v,u}^z$$

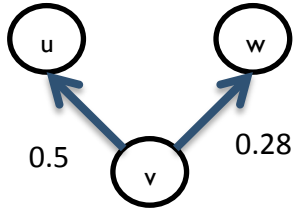
News




Economics




Sport



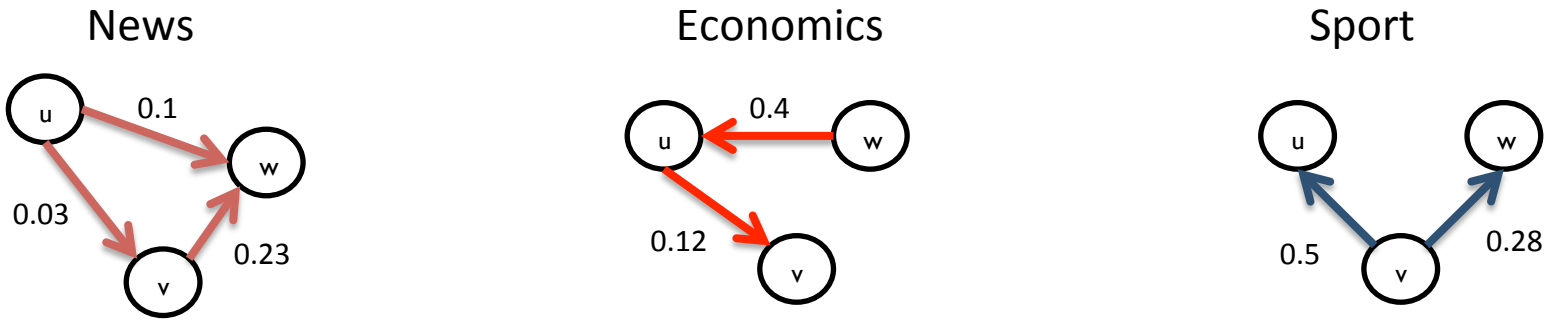
# Topic-Aware Independent Cascade (TIC)

**For each item that propagates in the network, we have a distribution over topics**

$$\gamma_i^z = P(Z = z|i)$$

**User-to-user influence probabilities depend on the topic**

$$p_{v,u}^i = \sum_{z=1}^{K} \gamma_i^z p_{v,u}^z$$



# Given:

Social Network

Propagation Cascades

We propose an EM algorithm to effectively **LEARN** the parameters of the TIC model

**E-step**

**forall the** $i \in \mathcal{I}$ **do**
  **forall the** $z = \{1, \cdots, K\}$ **do**
    $Q_i(z; \hat{\Theta}) \leftarrow \frac{P(D_i|z;\hat{\Theta})\pi_z}{\sum_{\tilde{z}} P(D_i|\tilde{z};\hat{\Theta})\pi_{\tilde{z}}}$;
    **forall the** $(u, v) \in E$ **do**
      $R_z^i(u, v; \hat{\Theta}) \leftarrow \frac{p_{v,u}^z}{P_{u,+}^{i,z}}$;
    **end**
  **end**
**end**

**M-step**

**forall the** $z = \{1, \cdots, K\}$ **do**
  $\pi_z \leftarrow \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} Q_i(z; \hat{\Theta})$;
  **forall the** $(u, v) \in E : S_{v,u}^+ \neq \emptyset$ **do**
    $p_{v,u}^z \leftarrow \frac{1}{\kappa_{v,u,z}^+ + \kappa_{v,u,z}^-} \sum_{i \in S_{v,u}^+} Q_i(z; \hat{\Theta}) R_z^i(u, v; \hat{\Theta})$
  **end**
**end**

# Good News:

The expected spread σ(S) remains monotone and submodular for TIC

The simple greedy algorithm for the Influence maximization problem provides an approximation guarantee

… but:

TIC is characterized by a huge number of parameters
#topics( #links + #items)!!!!

TIC assumes single influence. However, in many real-life scenarios propagation is the result of cumulative influence

# The AIR Propagation Model

**Authoritativeness** of a user in a topic: $p_v^z \in \mathbb{R}$

**Interest** of a user for a topic: $\vec{\vartheta}_u$

**Relevance** of an item for a topic: $\varphi_i^z \in \mathbb{R}$


Lady Gaga @ladygaga

Justin Bieber @justinbieber


Barack Obama @barackobama

CNN @cnn

The Economist @TheEconomist



0.92 0.08



0.01 0.09

# AIR: Activation Model

AIR is a **general threshold model**

Each user exhibits different degree of interest in different topics

$$P(i|u,t) = \sum_z P(z|u) P(i|u,z,t) \geq \theta_u$$

Likelihood of the activation on the item (i) when the topic is (z)

Item Selection Weight for the considered topic

Cumulative influence by neighbors

$$P(i|u,z,t) = \frac{\exp\left\{\sum_{v \in V} p_v^z f_v(i,u,t) + \varphi_i^z f(i,u,t)\right\}}{1 + \exp\left\{\sum_{v \in V} p_v^z f_v(i,u,t) + \varphi_i^z f(i,u,t)\right\}}$$

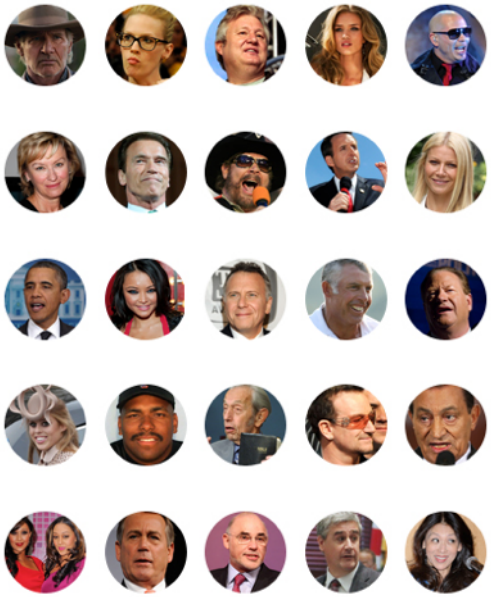
Selection scaling factors

# Given:

Social Network

Propagation Cascades

# We want to learn

# Parameter Estimation

Assuming that the hidden topic variable is independent from time, the **complete data log likelihood** is:

$$\mathcal{L}(\Theta; Q) = \sum_i \sum_u \sum_z Q(z; u, i) \{\log \vartheta_u^z + \sum_{\underline{t_i}}^{\overline{t_i}} d_i^u(t) \log P(i|u, z, t) + (1 - c_i^u(t)) \log (1 - P(i|u, z, t))\}$$

$D_i(t)$ denotes the set of users who selected the item $i$ at time $t$

$C_i(t)$ denotes the set of users who selected $i$ by time $t$

$d_i^u(t) = 1$ if $u \in D_i(t)$

$c_i^u(t) = 1$ if $u \in C_i(t)$

The non-linearity of the selection function makes it difficult to maximize the likelihood

**Solution**: ***Improved Iterative Scaling*** algorithm and the ***Generalized Expectation-Maximization*** procedure

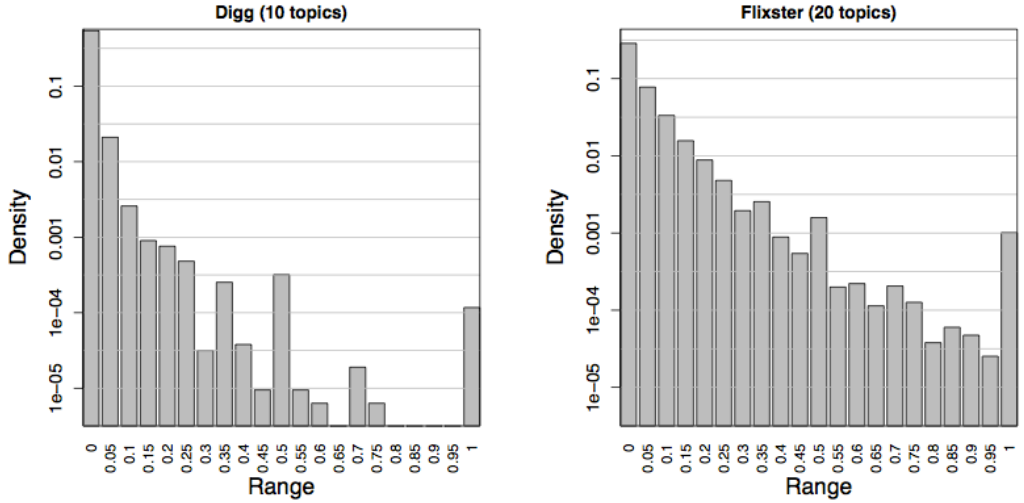
$$\mathcal{L}(\Theta + \Gamma, Q) \geq \mathcal{L}(\Theta, Q)$$

# Experimental Evaluation

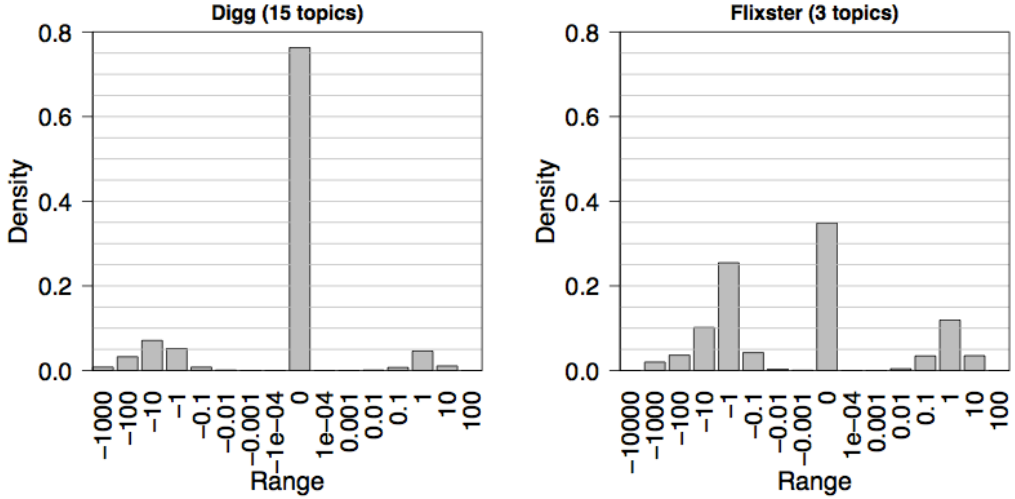
1. Assessing the capabilities of the Topic-Aware Propagation Models in **predicting users behavior**
2. Item Aware **Influence Maximization Problem**

| | FLIXSTER | | DIGG | |
|---|---|---|---|---|
| | Training | Test | Training | Test |
| Users | 6,572 | 4,686 | 16,297 | 14,061 |
| Items | 7,158 | 7,138 | 3,553 | 3,547 |
| Actions | 1,432,716 | 340,495 | 1,160,428 | 264,066 |
| Avg # actions (user) | 218 | 72 | 71 | 18 |
| Avg # actions (item) | 200 | 47 | 326 | 74 |
| Min # actions (user) | 6 | 1 | 6 | 1 |
| Min # actions (item) | 9 | 1 | 90 | 3 |
| Max # actions (user) | 5,525 | 1,786 | 2,640 | 1,912 |
| Max # actions (item) | 3,173 | 778 | 4,995 | 828 |
| Avg lifetime (item) | 952 days | | 14 days | |
| | Avg time between two actions | | | |
| per user | 94 hours | | 66 hours | |
| per item | 22 days | | 38 minutes | |

# Analyzing the Distributions of the influence weights




Distribution of $p^z_{v,u}$ in the TIC model




Distribution of $p^z_v$ in the AIR model.

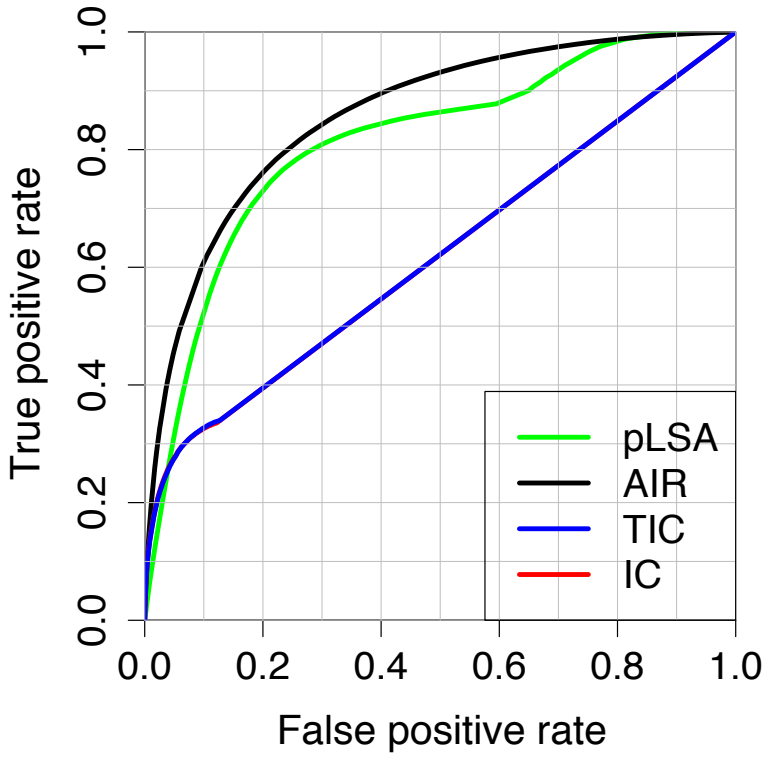
Influence probabilities are **exponentially distributed**

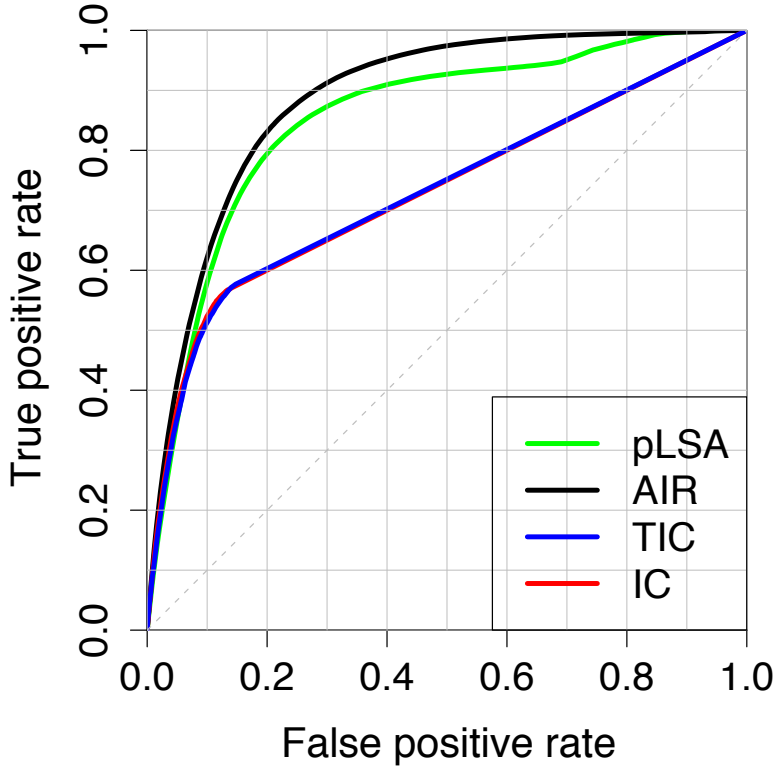
The Digg dataset exhibits a lower level of influence among users

Values are distributed according to **two log-normal distributions** centered in the positive and negative quadrants

# Predictive Accuracy: Activation Test

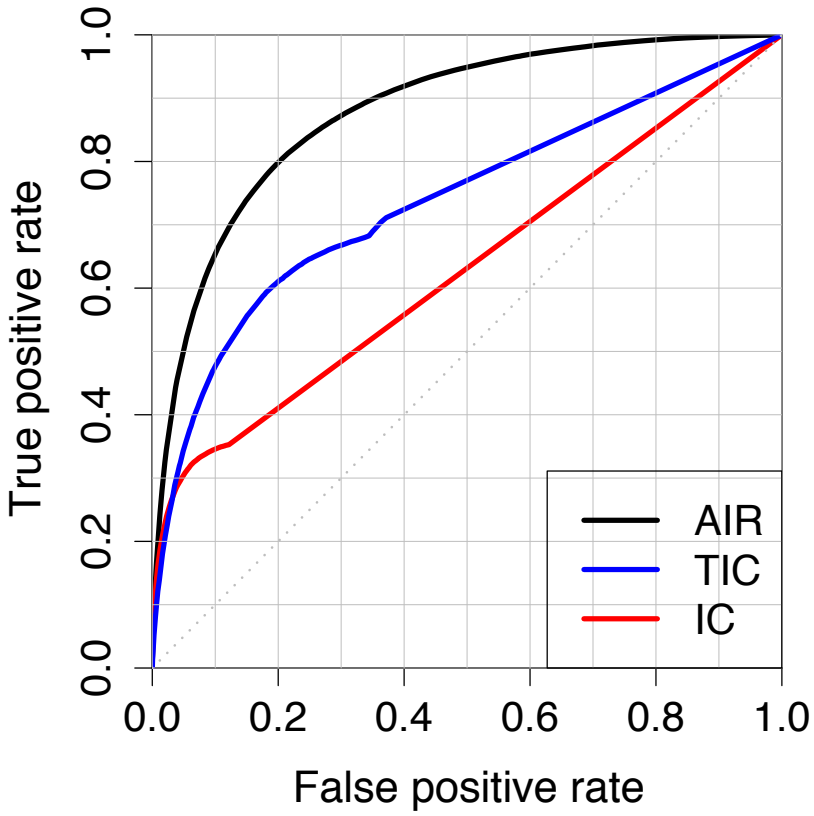
For any user-item pair ⟨u,i⟩ not observed in the training, we try to predict whether it belongs to the test

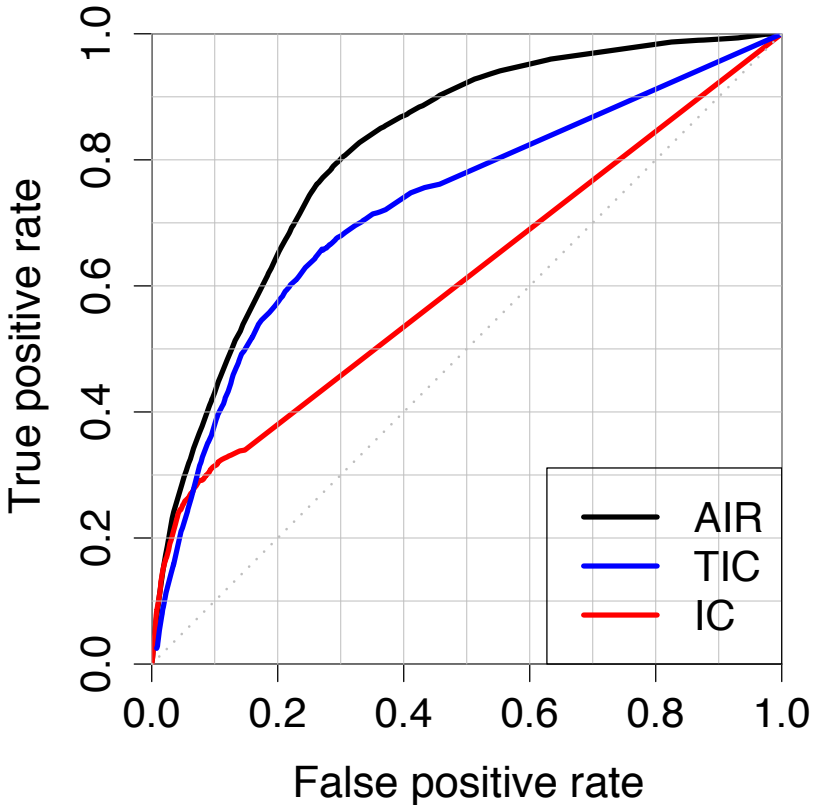


# Predictive Accuracy: Selection Probability

For any user-item pair ⟨u, i⟩ not observed in the training, such that the set of potential influencers is not empty, we measure the degree of responsiveness of the model at the actual activation time $t_i(u)$ (if it exists)

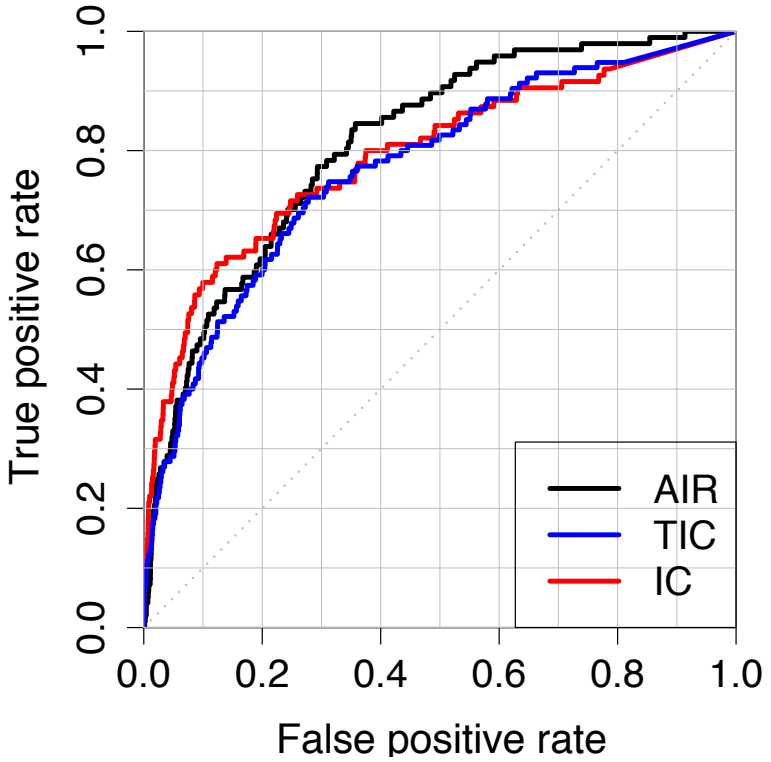


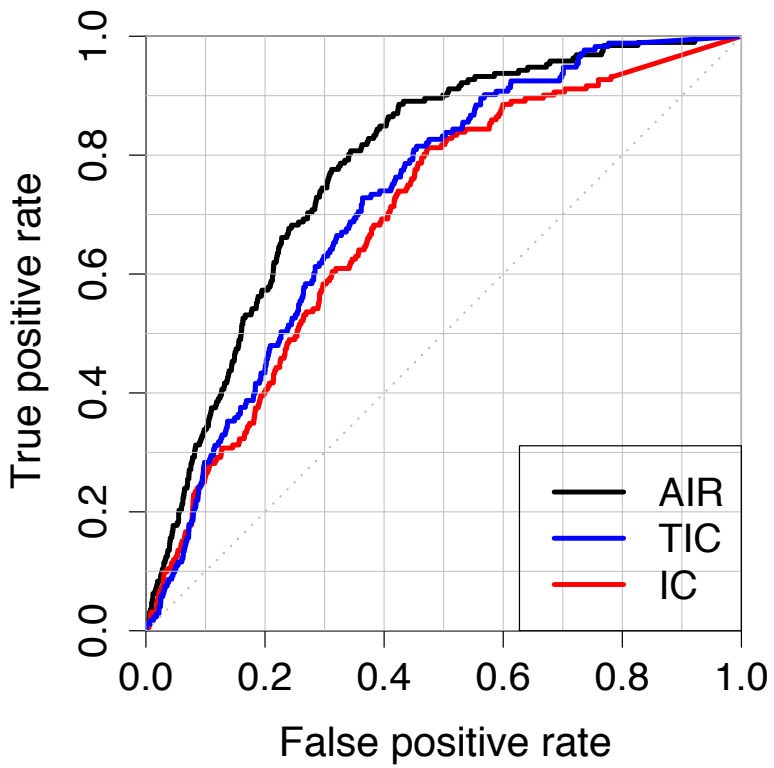



# Predictive accuracy: Activation time

For each pair ⟨u,i⟩ not observed in the training is evaluated by comparing the true activation time (if any) with the predicted activation time.

| | $\langle u, i\rangle \in \mathbb{D}_{Test}$ | $\langle u, i\rangle \notin \mathbb{D}_{Test}$ |
|---|---|---|
| True Positive | $t'_i(u) = t_i(u)$ | - |
| False Positive | $t'_i(u) < t_i(u)$ | $t'_i(u) \neq \infty$ |
| True Negative | - | $t'_i(u) = \infty$ |
| False Negative | $t'_i(u) > t_i(u)$ | - |



# Predictive Accuracy: Summary

| Model | DIGG | FLIXSTER | DIGG | FLIXSTER |
|---|---|---|---|---|
| | Activation Test (General) | | Selection Probs. (General) | |
| AIR | **0.8585511** | **0.8857634** | **0.8484368** | **0.8201586** |
| TIC | 0.6190136 | 0.731208 | 0.6256339 | 0.7000218 |
| IC | 0.6189209 | 0.730694 | 0.5256555 | 0.702175 |
| | Selection Probs. (Inf. episodes) | | Activation Time | |
| AIR | **0.8123432** | **0.783486**4 | **0.8784483** | **0.8150082** |
| TIC | 0.7714797 | 0.7253222 | 0.7377654 | 0.7377654 |
| IC | 0.7916101 | 0.6940882 | 0.6294611 | 0.6089646 |

AUC values

The AIR model achieves the best results in detecting the activations, with a consistent gain over the other models

When considering general activations TIC and IC exhibit partial ROC curves: negative samples are a vast majority, and this boost the number of True Negatives

Both TIC and AIR exhibit a consistent gain in predicting the activation time: this experimentally proves the effectiveness of the Topic-Aware techniques in modeling users' behavior

# Cascade-based Community Detection

Barbieri, Bonchi, Manco (WSDM'13)

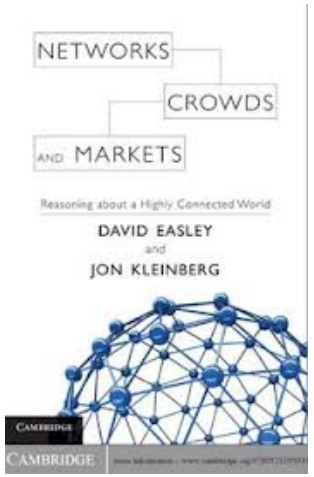
# Idea: to model the modular structure of SN and the phenomenon of social contagion *jointly*

Input: directed social graph + set of past propagations over the graph

Output: overlapping communities of nodes, *that also explain the cascades*.

Easley and Kleinberg book [page 577]

*"…cascades and clusters truly are natural opposites: clusters block the spread of cascades, and whenever a cascade comes to a stop, there's a cluster that can be used to explain why."*



# How: by fitting a unique stochastic generative model to the observed social graph and propagations

(think about Twitter as an example)

## assumption:

each observed action

forming a link (following somebody), tweeting (original content), re-tweeting

is the result of a stochastic process

## observations:

one user belongs to multiple topics/communities of interest

with different levels of active/passive involvement

a link usually can be explained by one and only one community

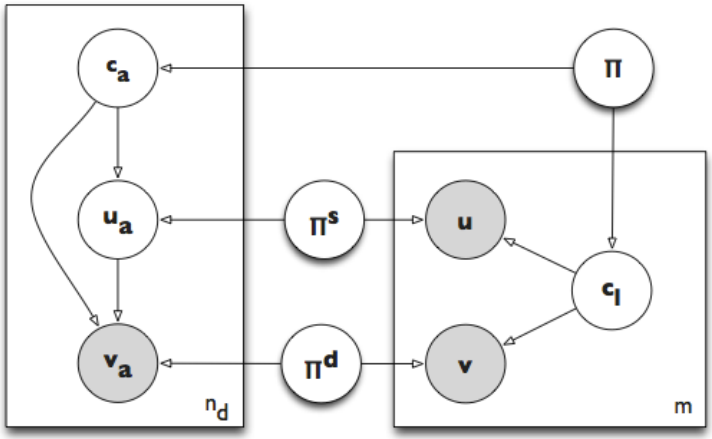
If I'm actively involved in a community I'm followed, I follow, I tweet and I re-tweet

If I'm passively involved in a community, I follow, I re-tweet,

but I'm not followed nor I tweet new content

# The CCN Model

(communities, cascades, network)




3 prior components:

the probability $\Pi$ to observe an action in a topic/community

the level of active $\Pi^s$ and passive $\Pi^d$ interest of each user in each community

each observed action is explained by the 3 priors

# Model parameters

- Probability of an edge through source/ destination

$$\vartheta_u^k = \frac{\exp\left\{\pi_u^{k,s}\right\}}{\sum_{\overline{u}\in N}\exp\left\{\pi_{\overline{u}}^{k,s}\right\}}$$

$$\varphi_u^k = \frac{\exp\left\{\pi_u^{k,d}\right\}}{\sum_{\overline{u}\in N}\exp\left\{\pi_{\overline{u}}^{k,d}\right\}}$$

- Probability of an action through influencer/ susceptible

$$\theta_u^{k,a} = \frac{\exp\left\{\pi_u^{k,s}\right\}}{\sum_{u'\in\mathcal{F}_{i_a}(t_a)}\exp\left\{\pi_{u'}^{k,s}\right\}}$$

$$\phi_{u,v}^{k,a} = \frac{\exp\left\{\pi_v^{k,d}\right\}}{\sum_{v':(u,v')\in A, v'\notin C_{i_a}(t_a-1)}\exp\left\{\pi_{v'}^{k,d}\right\}}$$

# Learning

- Expected likelihood not solvable analytically

$$Q(\Theta, \Theta') = \sum_{(u,v)\in A} \sum_{k} \gamma_{u,v,k}(\Theta') \left[\log \pi_k + \log \vartheta_u^k + \log \varphi_v^k\right]$$
$$+ \sum_{a\in\mathbb{D}} \sum_{k} \sum_{u\in\mathcal{F}_{i_a,v_a}} \eta_{u,a,k}(\Theta') \left(\log \pi_k + \log \theta_u^{k,a} + \log \phi_{u,v_a}^{k,a}\right)$$

- Resort to GEM again

# Learning (2)

E step

$$\gamma_{u,v,k}(\Theta) = P(z_\ell^k | \ell \equiv (u,v) \in A, \Theta)$$
$$= \frac{\vartheta_u^k \varphi_v^k \pi_k}{\sum_{k'} \vartheta_u^{k'} \varphi_v^{k'} \pi_{k'}}$$

$$\eta_{u,a,k}(\Theta) = P(z_a^k, w_a^u | a \in \mathbb{D}, \Theta)$$
$$= \frac{P(a \in \mathbb{D} | w_a^u, z_a^k, \Theta) P(w_a^u | z_a^k, \Theta) P(z_a^k | \Theta)}{P(a \in \mathbb{D} | \Theta)}$$
$$= \frac{\phi_{u,v_a}^{k,a} \theta_u^{k,a} \pi_k}{\sum_{k'} \sum_{u' \in \mathcal{F}_{i_a,v_a}} \phi_{u',v_a}^{k',a} \theta_{u'}^{k',a} \pi_{k'}}$$

# Learning (3)

- Find an update $\Gamma$ such that

$$Q(\Theta^{old} + \Gamma, \Theta^{old}) \geq Q(\Theta^{old}, \Theta^{old})$$

- Analytic solution:

$$\delta_{\bar{u}}^{k} = \log \left\{ \frac{\sum_{v:(\bar{u},v)\in A} \gamma_{\bar{u},v,k} + \sum_{a\in\mathbb{D}:\bar{u}\in\mathcal{F}_{i_a,v_a}} \eta_{\bar{u},a,k}}{\vartheta_{\bar{u}}^{k} \sum_{(u,v)\in A} \gamma_{u,v,k} + \sum_{a\in\mathbb{D}:\bar{u}\in\mathcal{F}_{i_a}(t_a)} \theta_{\bar{u}}^{k,a} \sum_{u\in\mathcal{F}_{i_a,v_a}} \eta_{u,a,k}} \right\}$$

$$\lambda_{\bar{v}}^{k} = \log \left\{ \frac{\sum_{u:(u,\bar{v})\in A} \gamma_{u,\bar{v},k} + \sum_{a:v_a=\bar{v}} \sum_{u\in\mathcal{F}_{i_a,\bar{v}}} \eta_{u,a,k}}{\varphi_{\bar{v}}^{k} \sum_{(u,v)\in A} \gamma_{u,v,k} + \sum_{a:\bar{v}\notin C_{i_a}(t_a-1)} \sum_{u\in\mathcal{F}_{i_a,v_a}:(u,\bar{v})\in A} \eta_{u,a,k}\phi_{u,\bar{v}}^{k,a}} \right\}$$

# Evaluation: Twitter

- Data collected on august 2012. **Edges** represent following relationships, while **actions** are hashtags or url's.

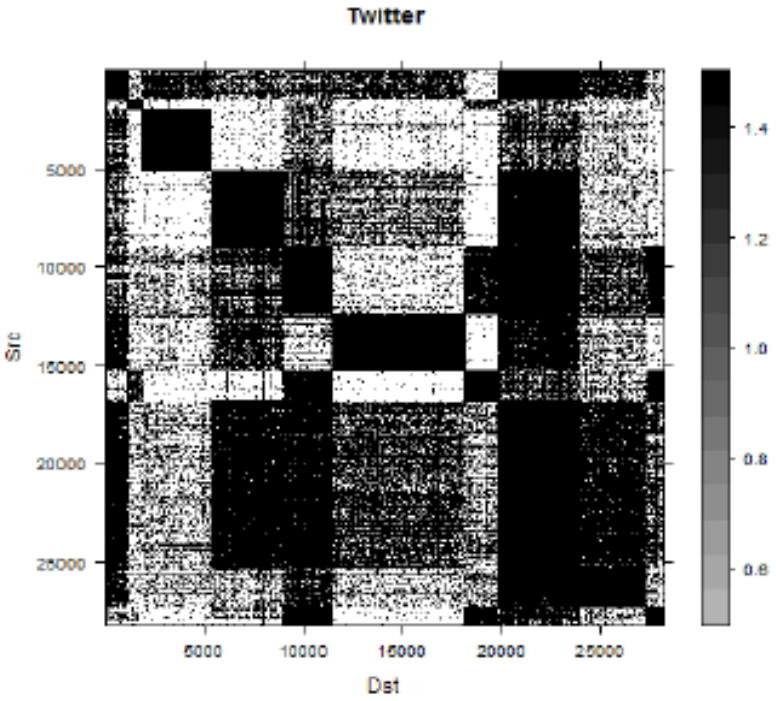
| | |
|---|---|
| **Users** | 32.042 |
| **Edges** | 1.636.451 |
| **Items** | 8.888 |
| **Actions** | 580.141 |

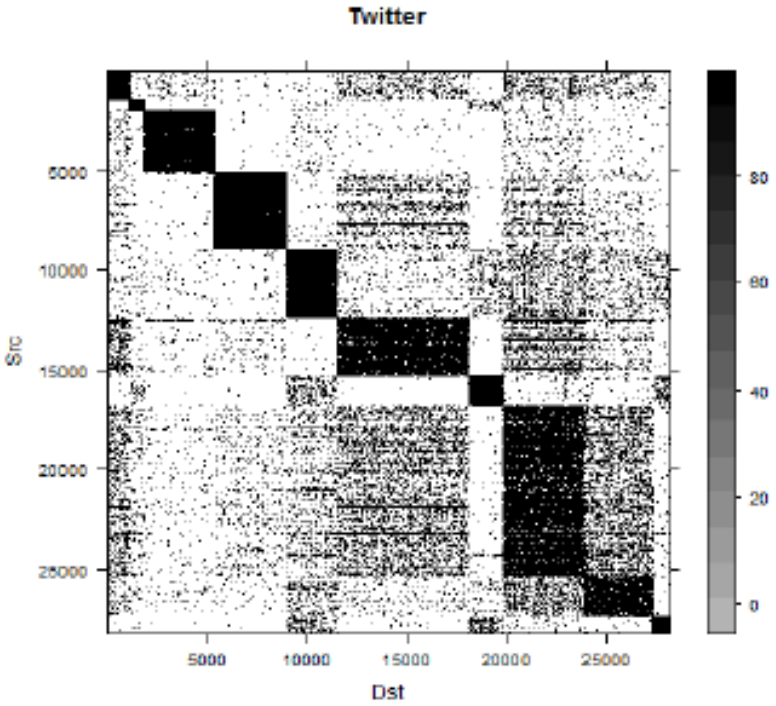
# Evaluation: Twitter

| | 2 | 4 | 8 | 16 | 32 | 64 |
|---|---|---|---|---|---|---|
| Likelihood ratio | 1.65E+06 | 3.35E+6 | 4.60E+6 | 5.08E+6 | 5.24E+6 | 4.79E+6 |
| Penalized log likelihood | -6.38E+07 | -6.26E+07 | -6.24E+07 | -6.39E+07 | -6.78E+07 | -7.61E+07 |
| Learning Time (hours) | 13.88 | 18.74 | 28.19 | 51.55 | 81.46 | 150.59 |
| Qg/ Qd (soft) | 1.49E-10/ 2.49E-03 | 2.11E-10/ 3.69E-03 | 2.57E-10/ 4.72E-03 | 2.71E-10/ 5.24E-03 | 2.89E-10/ 5.45E-03 | 2.86E-10/ 5.59E-03 |
| Qg/ Qd (hard) | 1.49E-10/ 2.23E-03 | 2.11E-10/ 3.03E-03 | 2.57E-10/ 3.46E-03 | 2.71E-10/ 3.60E-03 | 2.89E-10/ 3.43E-03 | 2.86E-10/ 3.27E-03 |

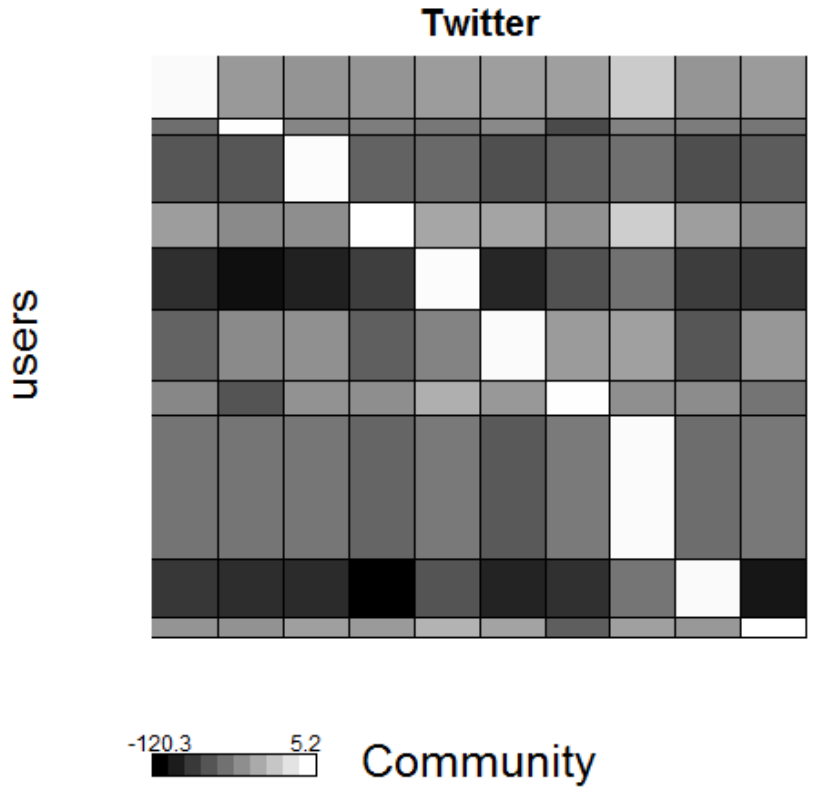
# Evaluation: Twitter

- 10 communities
- Adjacency/influence matrices

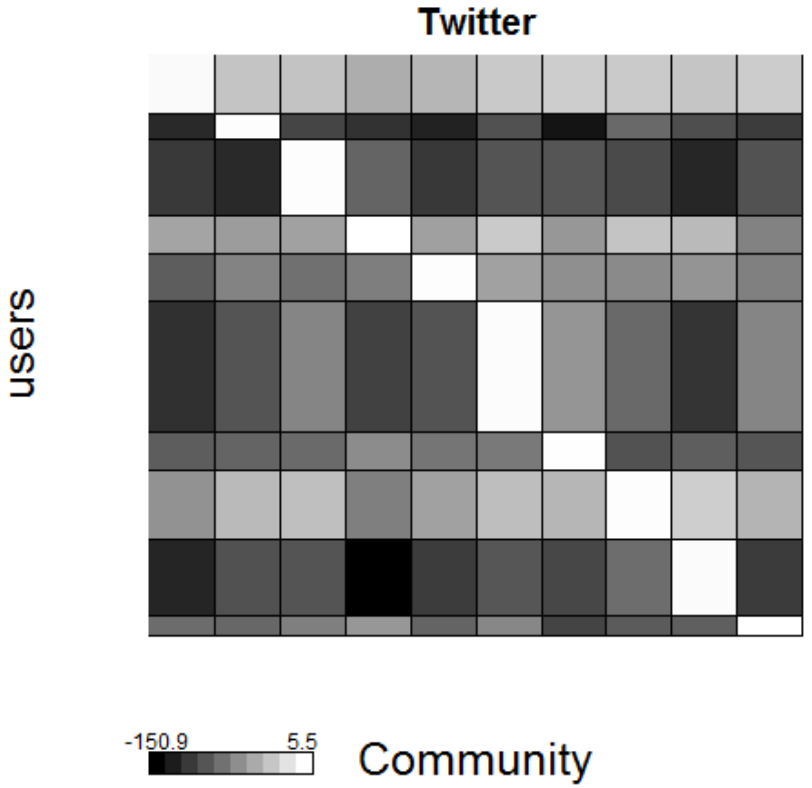


# Parameter distribution




Distribuzione di $\Pi\uparrow s$

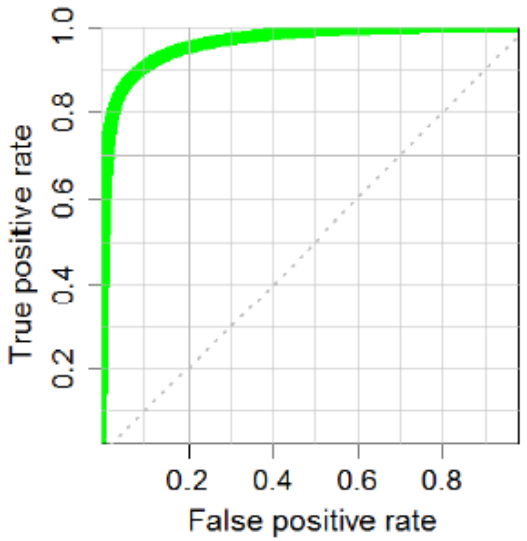


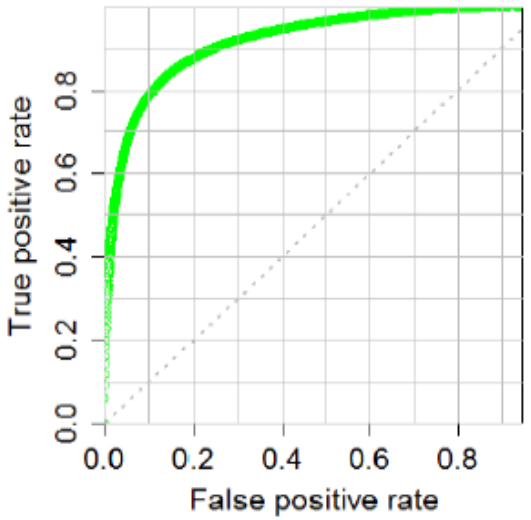

Distribuzione di $\Pi\uparrow D$

# Predictive abilities

- Link prediction:
  - How accurate is the model in predicting actual following relationships?

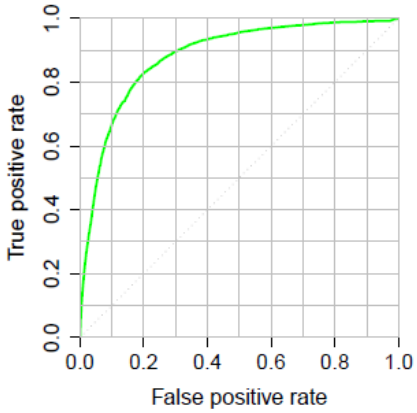
## Test Twitter e Hep-Ph (only network model)



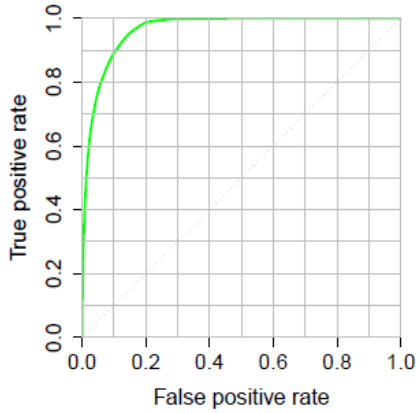



# Predictive tests on other datasets

## Test Digg (CCN e only network model)

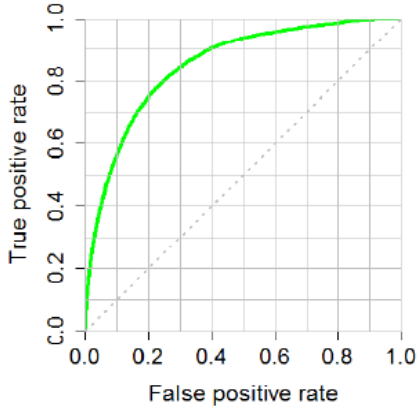



(a) Digg CCN model ROC curve

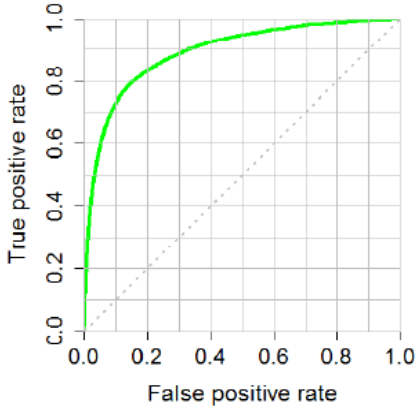



(b) LastFM bayesian model ROC curve.

## Test Last.fm (CCN e only network model)




(a) LastFM CCN model ROC curve.




(b) LastFM bayesian model ROC curve.

# Conclusions

- Strong probabilistic models
- Wide applicability: recommender systems, viral marketing, expert identification…
- Still missing: propagation rate
  - "virus" distribute ad different speeds
  - Can we reconstruct the social network by observing only their effects?

# THANKS!