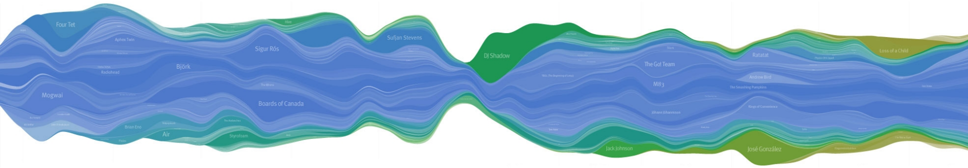
Information Spreading in Last.fm Online Social Network

Róbert Pálovics, András Benczúr

Informatics Laboratory, Department of Computer and Automation Research Institute, Hungarian Academy of Sciences



Supported by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE No 288956)

June 14, 2013

Outline

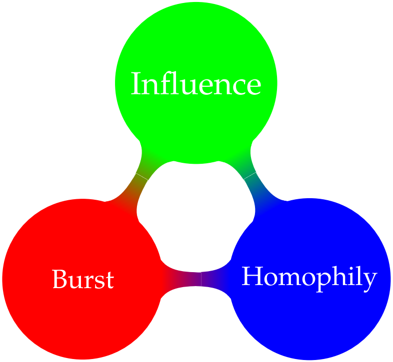
- Introduction
  - Last.fm
- Experimental results
  - Densification law
  - Temporal influences
- Influence based recommendation
  - Influence recommender
  - Baseline recommenders
- Influence recommender experiments
- Summary
  - Future plans

General problem, task

- Characterize *information diffusion*, or *information spreading* by investigating online social networks
- Create an online, social network based recommendation system

Social effects

- *Social influence*: Action of individuals induce their friends to act in a similar way
- *Homophily*: The tendency of individuals to associate and bond with similar others
- *Burst*: Herding, following the crowd



- N. Christakis and J. Fowler, "The spread of obesity in a large social network over 32 years," *New England Journal of Medicine*, 357(4):370–379, 2007.
- M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," in *Annual Review of Sociology*, 27:415–444, 2001.
- A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *WSDM*, pp. 241–250, ACM, 2010.
- F. Bonchi, "Influence propagation in social networks: A data mining perspective," *IEEE Intelligent Informatics Bulletin*, 12(1):8–16, 2011.

LAST.FM

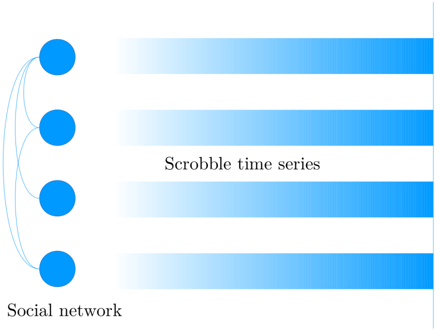
- About Last.fm
  - Leading online service in music based social networking
  - "Scrobbling": collecting listening activity of users
  - Recommendation system for users
  - Social network
- Influences
  - People often share their musical taste
  - They recommend each other new artists, albums, tracks
  - Directed influences

Dataset

- Available for us under NDA for Last.fm
- Selection criteria
  - User location is stated in UK
  - Age between 14 and 50, inclusive
  - Profile displays scrobbles publicly (privacy constraint)
  - Daily average activity between 5 and 500
- Size
  - 71,000 users, 285,241 edges
  - Scrobbles between 01 January 2010 and 31 December 2011 (2 year)
  - 979,391,001 scrobbles, 57,274,158 1st-time scrobbles
  - 2,073,395 artists

General task

- *User-user social network, with (scrobble) time series*
- Justify the existence of influences, i.e. correlation between individuals and the listening behavior of their contacts

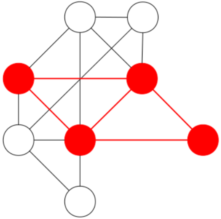


Artist subgraphs

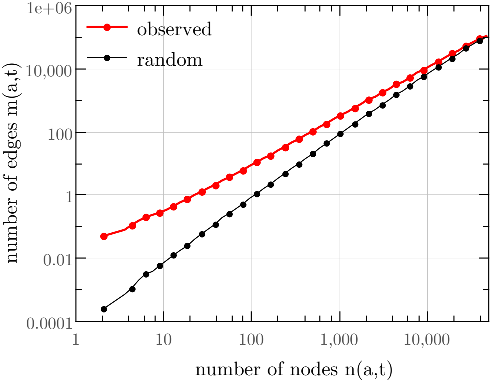
- For artist a in time t

$$G(a,t) = \{\text{subgraph of users who listened to } a \text{ before } t\}$$

- *Main result:*
  - **Increased edge density in** $G(a,t)$
  - The number of edges $m(a,t)$ is **power-law function** of the number of nodes $n(a,t)$ in the subgraph with exponent ≈ 1.535



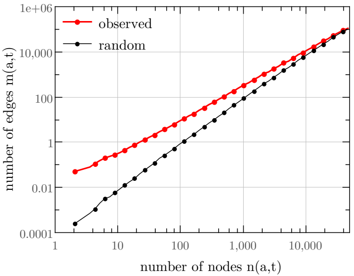
Measurements



- Larger graphs are denser
- But small artist subgraphs are much denser than random subgraphs

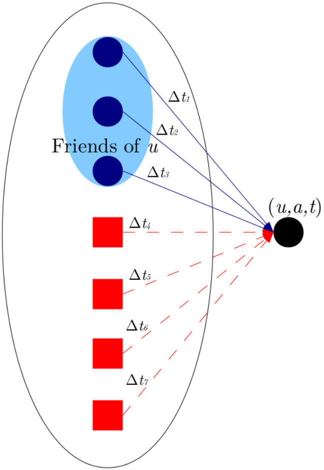
Future work

- Modeling densification law
- Analogies from statistical physics
- 2nd order phase transition (?)
- Problem: both endpoints refer to ordered states
- Finite size scaling (larger data → Twitter)



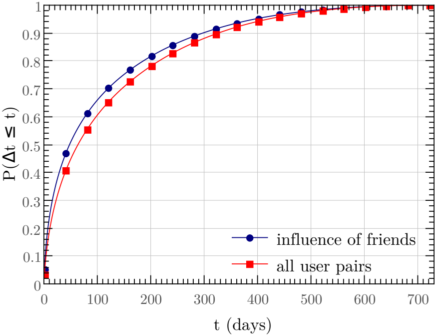
Temporal influence

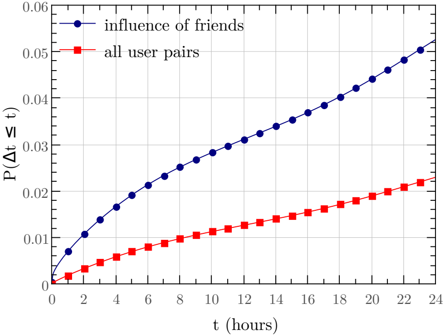
- User u is influenced by user v
- User u scrobbles a at the first time at t
- If v scrobbles a at time $t - \Delta t$
- Compute $\overline{\Delta t}$ in case of friends and all user pairs
- CDF(t) = fraction of influences with delay $\Delta t \leq t$ among all influences
- Friends vs. all pairs



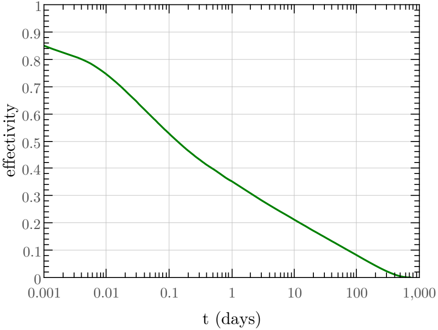
Users scrobbled a before t

CDF curves


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
100
200
300
400
500
600
700
P(Δt ≤ t)
t (days)
influence of friends
all user pairs


0.06
0.05
0.04
0.03
0.02
0.01
0
0
2
4
6
8
10
12
14
16
18
20
22
24
P(Δt ≤ t)
t (hours)
influence of friends
all user pairs

Effectivity curve



$$\mathrm{Eff}(\Delta t) = \frac{\mathrm{CDF}_F(\Delta t) - \mathrm{CDF}_A(\Delta t)}{\mathrm{CDF}_F(\Delta t)} \sim \log(\Delta t)$$

- Others propose exponential decay:
- A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *WSDM*, pp. 241–250, ACM, 2010.

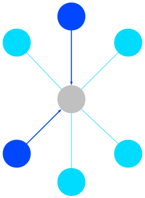
Recommender systems

- Predict the 'rating' or 'preference' that user would give to an item ($\hat{r}$)
- Top-k recommendation task: retrieve the best k items for the user u in time t
  1. Compute $\hat{r}(u, a, t)$ for all artists
  2. Order the artists
  3. Return the top-k elements in the list

a_1	$\hat{r}(a_1)$
a_2	$\hat{r}(a_2)$
a_3	$\hat{r}(a_3)$
a_4	$\hat{r}(a_4)$
a_5	$\hat{r}(a_5)$
a_6	$\hat{r}(a_6)$

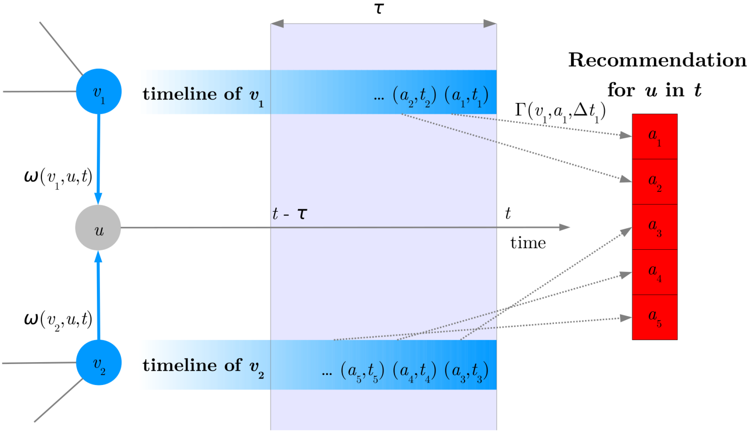
Main ideas

- Recommend artists scrobbled by her friends in the recent past
- Monotonically decreasing (logarithmic) dependence on time: $\Gamma(\Delta t(v, u, a))$
- Dependence of observed influence in the past: $\omega(v, u, t)$
- Score is the product of the two, for all friends



$$\hat{r}(u, a, t) = \sum_{v \in n(u)} \Gamma(\Delta t(v, u, a))\omega(v, u, t)$$

Influence recommender



$$\hat{r}(u,a,t) = \sum_{v \in n(u)} \Gamma(\Delta t(v,u,a))\omega(v,u,t)$$

INFLUENCE RECOMMENDER

- Influence function:

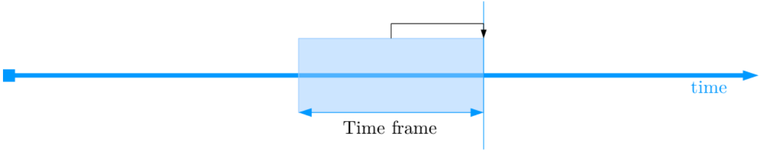
$$\Gamma(\Delta t(v, u, a)) = 1 - C \cdot \log(\Delta t),$$

- Strength between user pairs:
  1. $\omega(v, u, 0) = 0$
  2. $\omega(v, u, t_0) = \omega(u, v, t_0) = 1$
  3. $\omega(v, u, t) \leftarrow \omega(v, u, t) + (1 - C \cdot \log(\Delta t))$
- in case of time frame τ:

$$C = 1/\log \tau$$

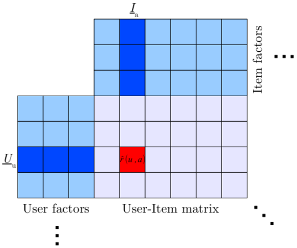
Dynamic popularity based recommendation

- Measure the popularity of an artist in $[t - \tau, t]$
- Recommend based on popularity scores
- Dynamic popularity based recommender $\Rightarrow$ global effects



Factor model based recommendation

- Factor model based recommenders became popular during the Netflix Prize competition[1]
- $\hat{r} = \underline{U}_u \cdot \underline{I}_a$
- A successful factor based recommender is described by Simon Funk[2]
- Optimize MSE by applying SGD method



[1] R. Bell and Y. Koren, “Lessons from the Netflix prize challenge,” 2007.
[2] “Netflix update: Try this at home http://sifter.org/šimon/journal/20061211.html,” 2006

FACTOR MODEL BASED RECOMMENDATION

- Iterate through the dataset
- At each record take a learning step
- Prediction: $\hat{r} = \underline{U}_u \cdot \underline{I}_a$
- Error: $\delta = r - \hat{r}$
- Objective function (with regularization rate α):

$$F = \frac{1}{2}\delta^2 + \alpha \cdot \left(||\underline{U}_u||^2 + ||\underline{I}_a||^2 \right) =$$

$$= \frac{1}{2}\left(r - \underline{U}_u \cdot \underline{I}_a\right)^2 + \alpha \cdot \left(||\underline{U}_u||^2 + \cdot||\underline{I}_a||^2 \right)$$

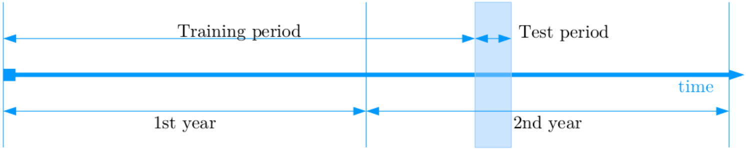
- Learning steps based on the gradient of F (learning rate: λ):

$$\underline{\Delta U}_u = \lambda \cdot \delta \cdot \underline{I}_a - \lambda \cdot \alpha \cdot \underline{U}_u$$

$$\underline{\Delta I}_a = \lambda \cdot \delta \cdot \underline{U}_u - \lambda \cdot \alpha \cdot \underline{I}_a$$

Factor model based recommendation

- Weekly trained models and computed top-k recommendations
- Train data: all scrobbles before the given week + negative scrobbles (3X)
- Factor model $\Rightarrow$ homophily



FUTURE WORK

- Present influence recommender:
  - heuristic weighted network learning ☹
  - no artist based learning part ☹
- Influence + factor model → learn how
  - likely influences user v with artist a user u
  - influencable is user u in case of artist a
- Use SGD method to learn user and artist factors

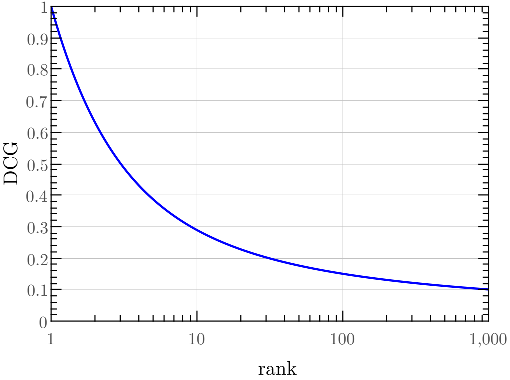
$$\hat{r}(u, a, t) = \frac{1}{deg(u)} \sum_{v \in n(u)} \Gamma(\Delta t(v, u, a)) \cdot (\underline{U}_v \cdot \underline{I}_a + \ldots)$$

Evaluation of top-k recommendation

- Influence scores rapidly change in time $\rightarrow$ separate evaluation for each individual scrobble
- Create a top-k list recommendation in case of each new user-artist scrobble (u, a, t)
- Measure the goodness of this returned list
- The lower is the rank of a in the returned list, the better is our prediction
- Discounted cumulative gain with threshold K

$$\mathrm{DCG@K}(a) = \begin{cases} 0 & \text{if rank } (a) > K; \\ \dfrac{1}{\log_2(\mathrm{rank}(a) + 1)} & \text{otherwise.} \end{cases}$$

EVALUATION OF TOP-k RECOMMENDATION

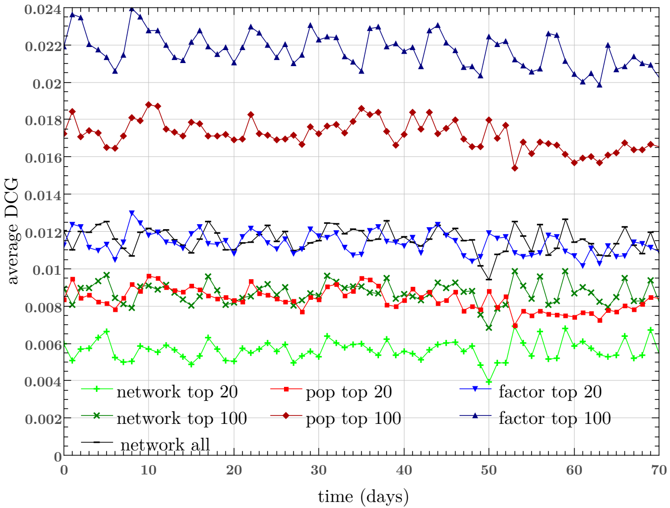


$$\mathrm{DCG@K}(a) = \begin{cases} 0 & \text{if rank } (a) > K; \\ \dfrac{1}{\log_2(\mathrm{rank}(a)+1)} & \text{otherwise.} \end{cases}$$

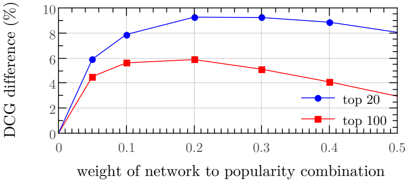
EVALUATION OF TOP-k RECOMMENDATION

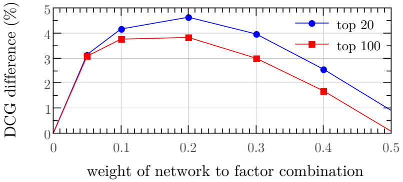
- Compute DCG@K score for all 1st-time scrobble in the 2nd year
- Compute time-averages over DCG@K scores
- Always use the 1st year as a training set
- Every recommender can use all scrobbles before the evaluated one as training data

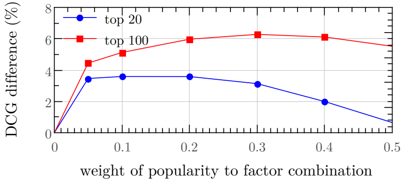
Results

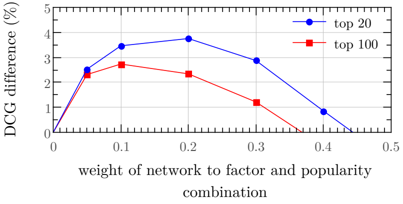

0.024
0.022
0.02
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
average DCG
0
10
20
30
40
50
60
70
time (days)
network top 20
pop top 20
factor top 20
network top 100
pop top 100
factor top 100
network all

Combination

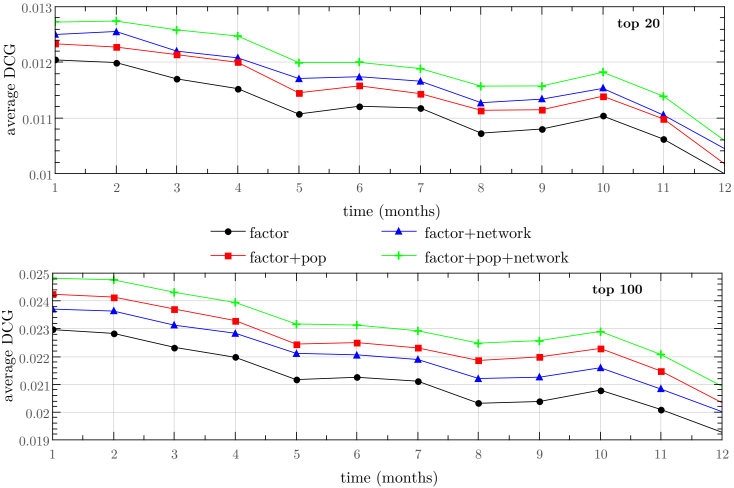








COMBINATION



CONCLUSIONS

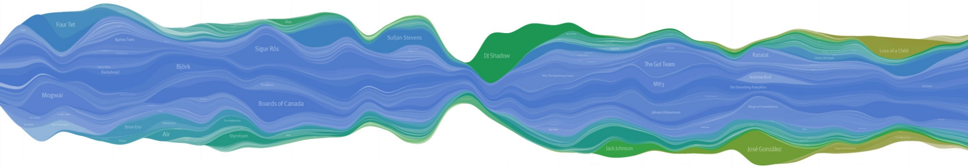
- 70,000 users, 979,391,001 scrobbles, 57,274,158 1st-time scrobbles
- Basic influence measurements (densification law, artist subgraphs)
- Influence based recommender system
- Lightweight, fast, easy to implement influence recommender

Conclusions

- Baseline recommenders that take homophily and global effects into account
- Strong, never vanishing improvement of baseline methods by combining them with influence based recommendation
- Results confirm the existence of social influence

Twitter

- Tweets, retweets, topics over a social network
- Evolution of one topic (e.g. #occupy, ...) $\Leftrightarrow$ evolution of a popular artist
- Set of retweets $\Leftrightarrow$ evolution of an artist
- In case of a retweet we only know the original *tweet source*(!)
- $\Rightarrow$ Last.fm measurements can be repeated with Twitter datasets
- Last.fm: influence pairs $\leftrightarrow$ Twitter: large retweet cascades
- Temporal evolution of retweet cascades


Four Tet
Aphex Twin
Radiohead
Mogwai
Björk
Brian Eno
Air
Sigur Rós
Boards of Canada
Styrofoam
Sufjan Stevens
DJ Shadow
The Go! Team
M83
Jack Johnson
Ratatat
Andrew Bird
The Smashing Pumpkins
José González
Loss of a Child