On social influence, topics, and communities

Francesco Bonchi

www.francescobonchi.com

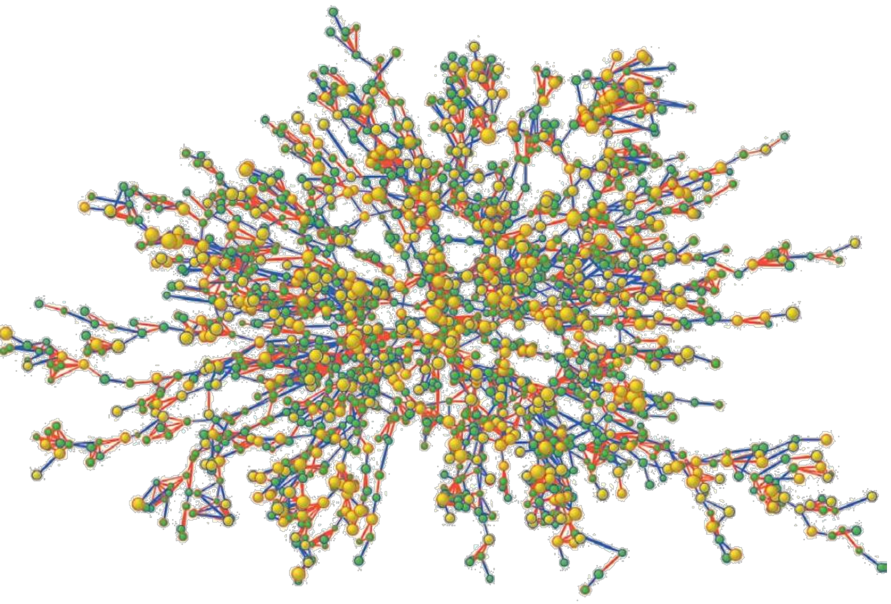
Plan of the talk

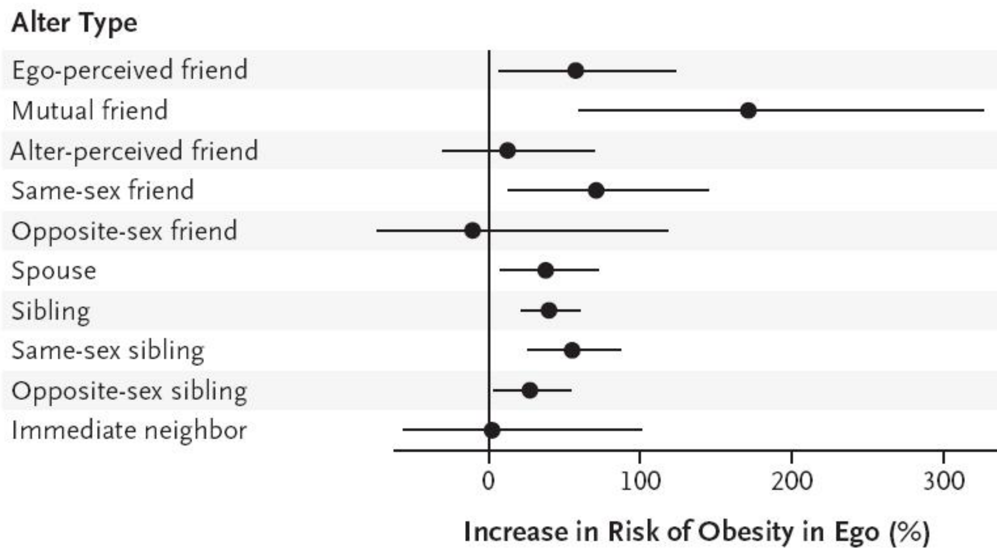
- Some background on social influence
- Some background on influence maximization
- Topic-aware social influence propagation models
- Cascade-based community detection
- Who to Follow and Why: Link Prediction with Explanations

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?

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

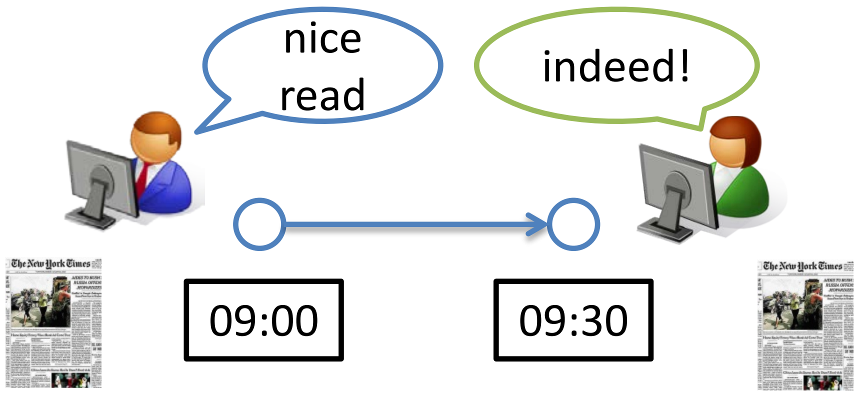
Influence is a causal process

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"*

On-going project: Developing computational methods for understanding social influence using **Suppe's Probabilistic Causation theory** [joint work with Bud Mishra and Daniele Ramazzotti].

Influence-driven information propagation 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

Mining propagation data: opportunities
(science, society, technology and business)

studies and models of human interaction

innovation adoption, epidemics

social influence, homophily, interest, trust, referral

citizens engagement, awareness, law enforcement

citizens journalism, blogging and microblogging

outbreak detection, risk communication, coordination during emergencies

political campaigns

feed ranking, personalization, expert finding, “friends” recommendation

branding

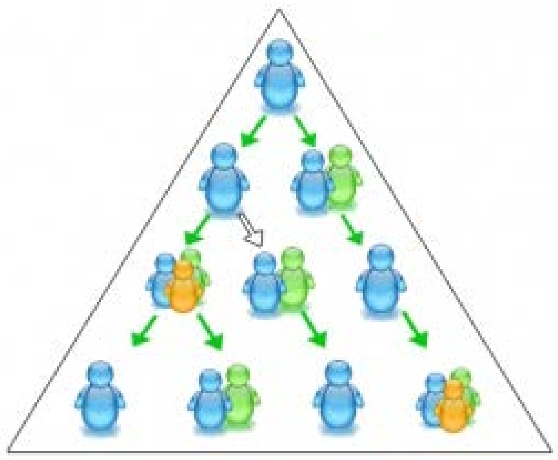
behavioral targeting

WOMM, viral marketing

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 14 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)

Influence Maximization Problem

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

Given a propagation model *M,* define influence of node set *S,*
$\sigma_M(S)$ = expected size of propagation, if *S* is the initial set of active nodes

Problem: Given social network *G* with arcs probabilities/weights, budget k, find k-node set *S* that maximizes $\sigma_M(S)$

Two major propagation models considered:

independent cascade (IC) model

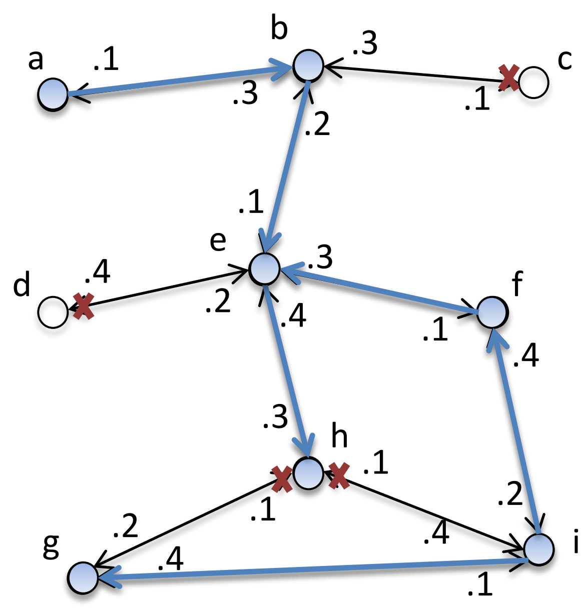
linear threshold (LT) model

Independent Cascade Model (IC)

Every arc *(u,v)* has associated the probability *p(u,v)* of *u* influencing *v*

Time proceeds in discrete steps

At time t, nodes that became active at t-1 try to activate their inactive neighbors, and succeed according to *p(u,v)*



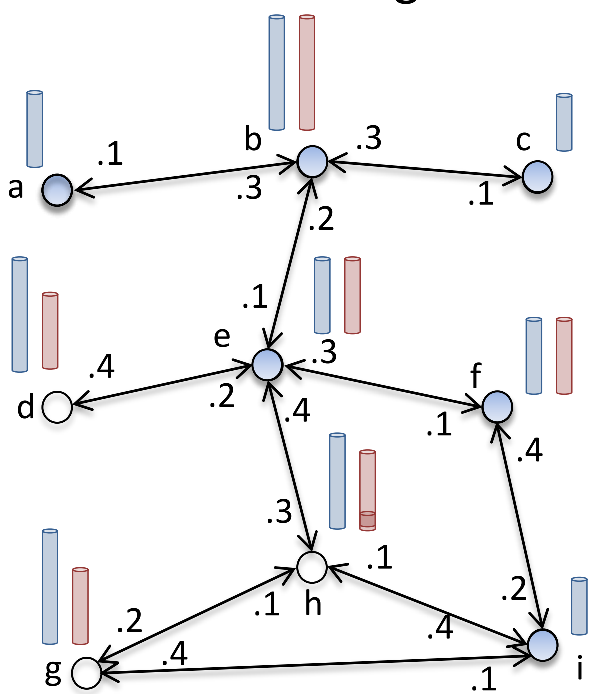
Linear Threshold Model (LT)

Every arc *(u,v)* has associated a weight *b(u,v)* such that the sum of incoming weights in each node is ≤ 1

Time proceeds in discrete steps

Each node *v* picks a random threshold $\vartheta_v \sim U[0,1]$

A node *v* becomes active when the sum of incoming weights from active neighbors reaches ϑ_v



Known Results

Bad news: NP-hard optimization problem for both IC and LT models

Good news: we can use Greedy algorithm

Algorithm 1 Greedy

Input: G, k, σ_m

Output: seed set S

1: $S \leftarrow \emptyset$

2: **while** $|S| < k$ **do**

3: select $u = \arg\max_{w \in V \setminus S}(\sigma_m(S \cup \{w\}) - \sigma_m(S))$

4: $S \leftarrow S \cup \{u\}$

$\sigma_M(S)$ is monotone and submodular

Theorem*: The resulting set S activates at least (1- 1/e) > 63% of the number of nodes that any size-k set could activate

Bad news: computing $\sigma_M(S)$ is #P-hard under both IC and LT models

step 3 of the Greedy Algorithm is approximated by MC simulations

*Nemhauser et al. "An analysis of approximations for maximizing submodular set functions – (i)" (1978)

Influence Maximization algorithms

Much work has been done following Kempe et al. mostly devoted to heuristichs to improve the efficiency of the Greedy algorithm:

E.g.,

Kimura and Saito (PKDD’06) *“Tractable models for information diffusion in social networks”*

Leskovec et al. (KDD'07) *“Cost-effective outbreak detection in networks”*

Chen et al. (KDD'09) *“Efficient influence maximization in social networks”*

Chen et al. (KDD'10) *“Scalable influence maximization for prevalent viral marketing in large-scale social networks”*

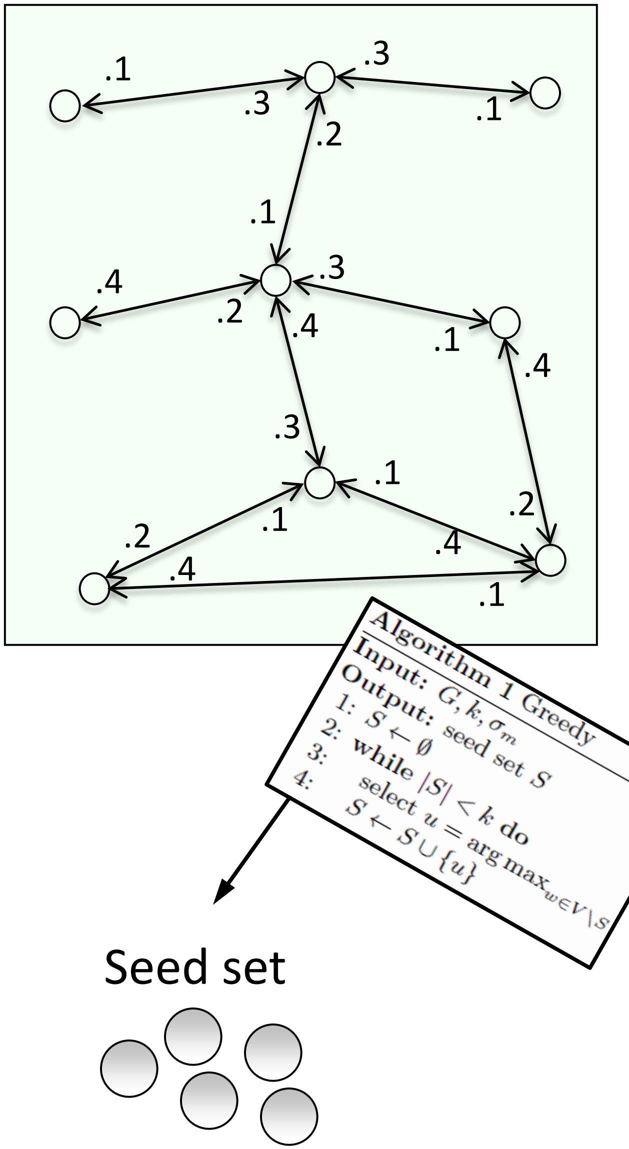
Goyal et al. (WWW’11) *“CELF++: optimizing the greedy algorithm for influence maximization in social networks”*

… … …

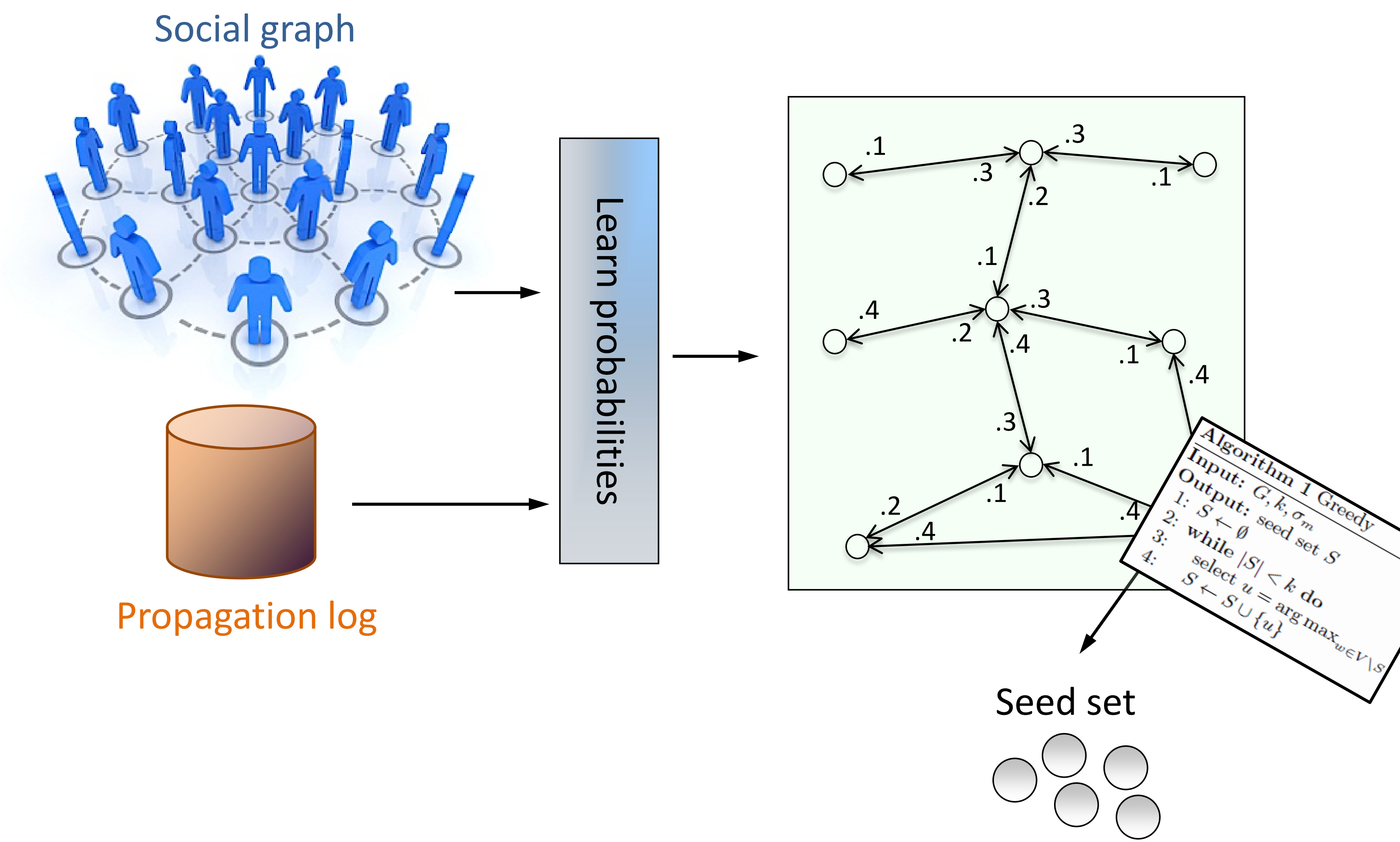
Borgs et al. (SODA’14) *“Maximizing social influence in nearly optimal time”*

Tang et al. (SIGMOD’14) *“Influence maximization: Near-optimal time complexity meets practical efficiency”*

Cohen et al. (CIKM’14) *“Sketch-based influence maximization and computation: Scaling up with guarantees”*



The larger picture of Influence Maximization



Data! Data! Data!

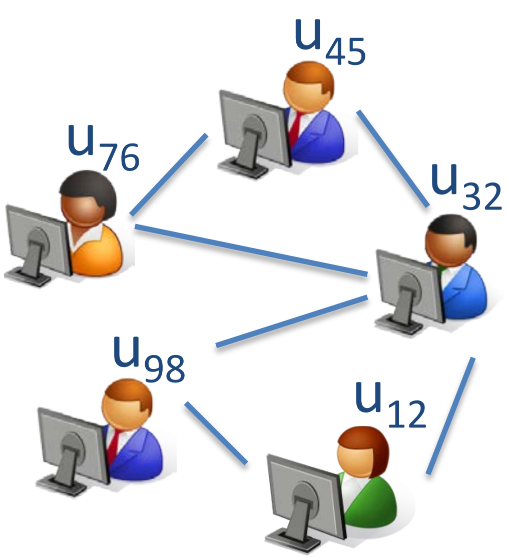
We have 2 pieces of input data:
(1) social graph and (2) a log of past propagations

Putting together (1) and (2) we can consider to have

a set of **DAGs**

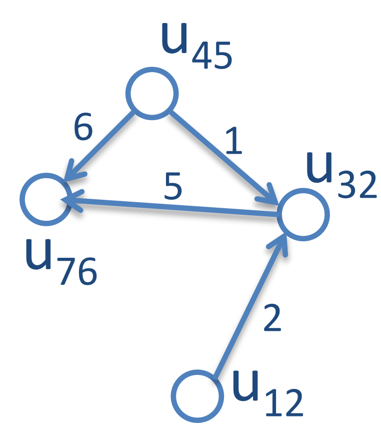
(sometimes a set of trees)

with arcs labeled with elapsed time between two actions



Action	User	Time
a	u_{12}	1
a	u_{45}	2
a	u_{32}	3
a	u_{76}	8
b	u_{32}	1
b	u_{45}	3
b	u_{98}	7

Action *a:*



Learning influence strenght

A. Goyal, F. Bonchi, L. V. S. Lakshmanan

Learning Influence Probabilities In Social Networks (WSDM 2010)

N. Barbieri, F. Bonchi, G. Manco

Topic-aware Social Influence Propagation Models (ICDM 2012) (KAIS)

K. Kutzkov, A. Bifet, F. Bonchi, A. Gionis

STRIP: Stream Learning of Influence Probabilities (KDD 2013)

T. Tassa, F. Bonchi

Privacy Preserving Estimation of Social Influence (EDBT 2014)

Privacy-preserving learning of influence strength
(Tassa & Bonchi – EDBT'14)

Provider *P1*

propagation log L1

host H

social graph G

Provider *P2*

propagation log L2

How the 3 (or more) players can learn influence strength jointly without seeing each other data?

A typical Secure Multiparty Computation setting.

Topic-aware Social Influence Propagation Models

Nicola Barbieri, Francesco Bonchi, Giuseppe Manco
ICDM 2012, KAIS

Topic-aware Social Influence Propagation Models

(Barbieri, Bonchi, Manco ICDM'12)

The bulk of the literature on Influence Maximization is **topic-blind**:
the characteristics of the item being propagated are not considered
(it is just one abstract item)

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

Key observations:
users have different interests,
items have different characteristics,
similar items are likely to interest the same users.

Thus we take a topic-modeling perspective to jointly learn
items characteristics, users' interests and social influence.

Topic-aware Social Influence Propagation Models

(Barbieri, Bonchi, Manco ICDM'12)

We have *K* topics
for each item *i* that propagates in the network,
we have a distribution over the topics.
That is, for each topic $z \in [1, K]$
we have

$\gamma_i^z = P(Z = z|i)$ with $\sum_{z=1}^{K} \gamma_i^z = 1$

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)

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

Learning problem

Given the database of propagations, the social network, and an integer *K*
Learn the model parameters, i.e.,

γ_i^z and $p_{v,u}^z$

We devise an EM algorithm for 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

... but:
TIC has a huge number of parameters
#topics(#links + #items)

The AIR propagation model

Authoritativeness of a user w.r.t. a topic
Interest of a user for a topic
Relevance of an item for a topic

$$P(i|u,t) = \sum_z \overbrace{P(z|u)}^{\text{Each user exhibits different degree of interest in different topics}} \underbrace{P(i|u,z,t)}_{\text{Likelihood of the activation on the item (i) when the topic is (z)}} \geq \theta_u$$

$$P(i|u,z,t) = \frac{\exp\left\{\overbrace{\sum_{v \in V} p_v^z f_v(i,u,t)}^{\text{Cumulative influence by neighbors}} + \overbrace{\varphi_i^z f(i,u,t)}^{\text{Item Selection Weight for the considered topic}}\right\}}{1 + \exp\left\{\sum_{v \in V} p_v^z \underbrace{f_v(i,u,t)}_{} + \varphi_i^z \underbrace{f(i,u,t)}_{}\right\}}$$

Selection scaling factors

[Learning the model parameters: see paper (!)]

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)

Another way to cut down the number of parameters

From user-to-user influence analysis

to …

Community-level Social Influence analysis

Network structure evolution, communities, cascades

N. Barbieri, F. Bonchi, G. Manco

Cascade-based Community Detection (WSDM 2013)

L. Weng, J. Ratkiewicz, N. Perra, B. Gonçalves, C. Castillo,

F. Bonchi, R. Schifanella, F. Menczer, A. Flammini

The Role of Information Diffusion in the Evolution of Social Networks (KDD 2013)

Y. Mehmood, N. Barbieri, F. Bonchi, A. Ukkonen

CSI: Community-level Social Influence analysis (ECML/PKDD 2013)

N. Barbieri, F. Bonchi, G. Manco

Influence-based Network-oblivious Community Detection (ICDM 2013)

N. Barbieri, F. Bonchi, G. Manco

Who to Follow and Why: Link Prediction with Explanations (KDD 2014)

Cascade-based Community Detection

Nicola Barbieri, Francesco Bonchi, Giuseppe Manco

WSDM 2013

State of the art

Social contagion

- measuring social influence
- distinguishing social influence from homophily in the data
- analysis of influence-driven information propagation in social media
- influence maximization

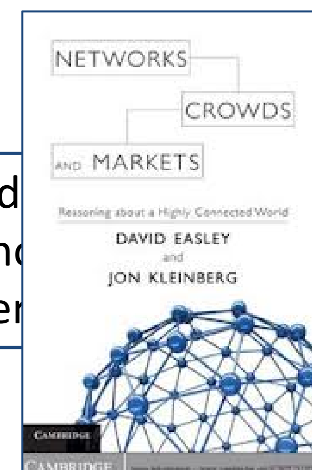
?

Community detection

- undirected Vs. directed graphs
- disjoint Vs. overlapping communities
- unlabeled Vs. labeled graphs

> *"...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."*

Easley and Kleinberg book [page 577]



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

Input:

directed social graph + a DB of past propagations over the graph

arc *(u,v)* means that *v* “follows” *u*

the DB of propagations is a set of tuples *(i,u,t)*

representing the fact that *u* adopted *i* at time *t*

Output:

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

for each node we also learn the level of

active involvement (i.e., tendency to produce content)

and passive involvement (i.e., tendency to consume content)

in each community

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

assumption:

each observed action

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

is the result of a stochastic process

observations:

(think about Twitter as an example)

one user belongs to multiple topics/communities of interest

with different levels of active/passive involvement

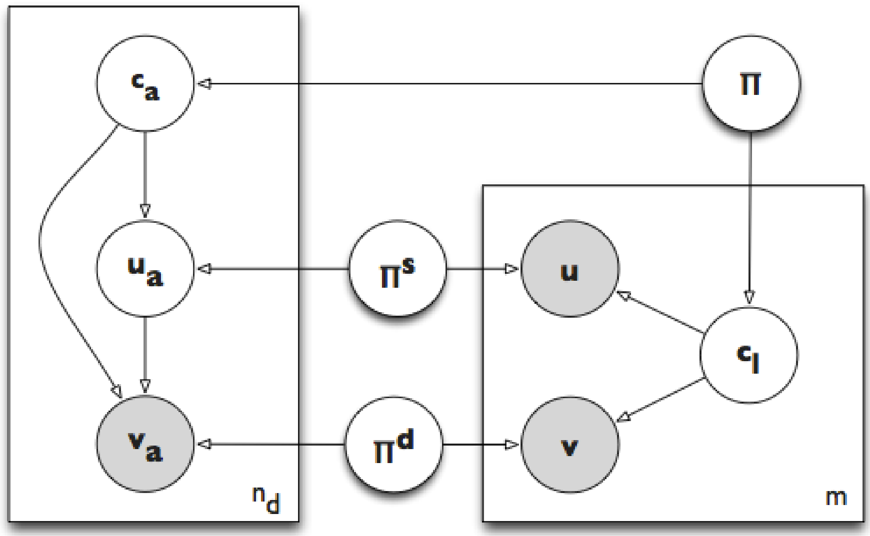
a link usually can be explained by one and only one community

If I'm actively involved in a community I'm followed, and I 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 Π to observe an action in a community
the level of active Π^s and passive Π^d interest of each user in each community

each observed action is explained by the 3 priors

The CCN Model (continued)

Probability of a link

(source)

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

(destination)

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

Probability of an action being propagated

(influencer)

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

(influenced)

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

Learning the model parameters

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

Solution adopted

Generalized Expectation-Maximization + Improved Iterative Scaling

(details in the paper!)

Experimental evaluation: datasets

	Digg	Flixster	Meme	LastFm
Users	1,000	29,357	9,385	1.372
Social Relationships	24,842	425,228	1,144,932	14,708
Bidirectional	N	Y	N	N
Items	31,911	11,659	12,760	51,495
Overall Activations($\|\mathbb{L}\|$)	1,086,065	6,529,011	726,809	1,208,640
Influence Episodes ($\|\mathbb{D}\|$)	315,377	2,239,744	684,368	322,932

Digg: social news website

Action *(i,u,t)* means that user *u* voted story *i* at time *t*

Flixster: social movie consumption (ranting and rating)

Action *(i,u,t)* means that user *u* rated movie *i* at time *t*

Meme (discontinued): microblogging platforms

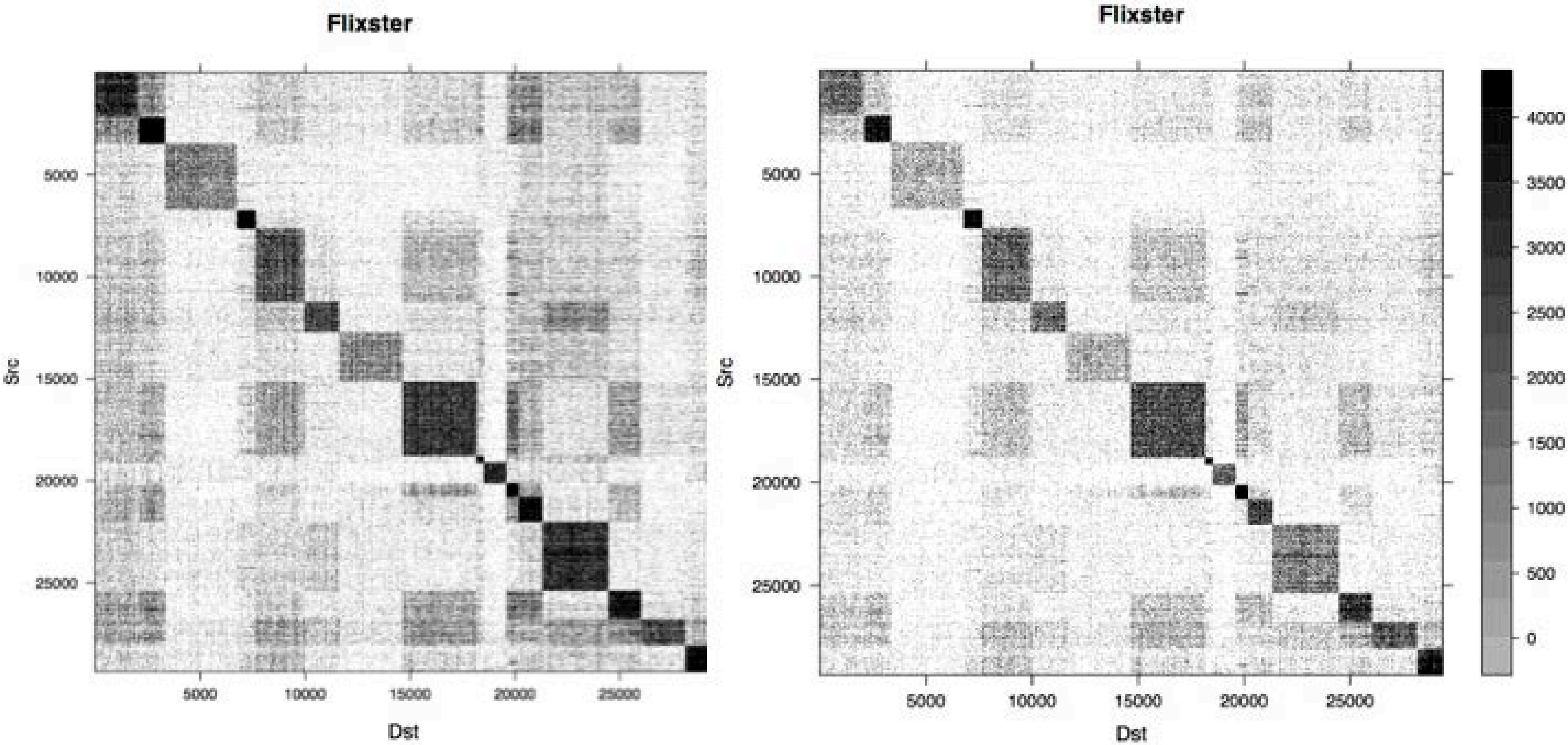
Action *(i,u,t)* means that user *u* posted meme *i* at time *t*

LastFM: social music consumption

Action *(i,u,t)* means that user *u* listened to song *i* at time *t*

Community structure within the graph and propagations DB

Adjacency matrix (left) and the influence matrix (right)
The influence matrix records for each cell *(u,v)* the number of actions for which the model infers that *u* triggered *v*'s activation

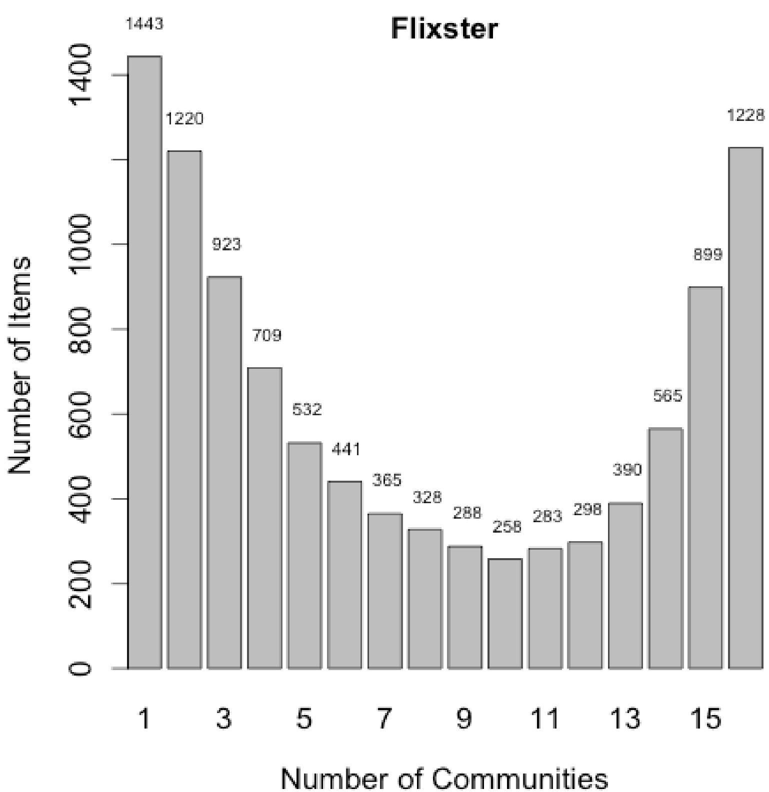


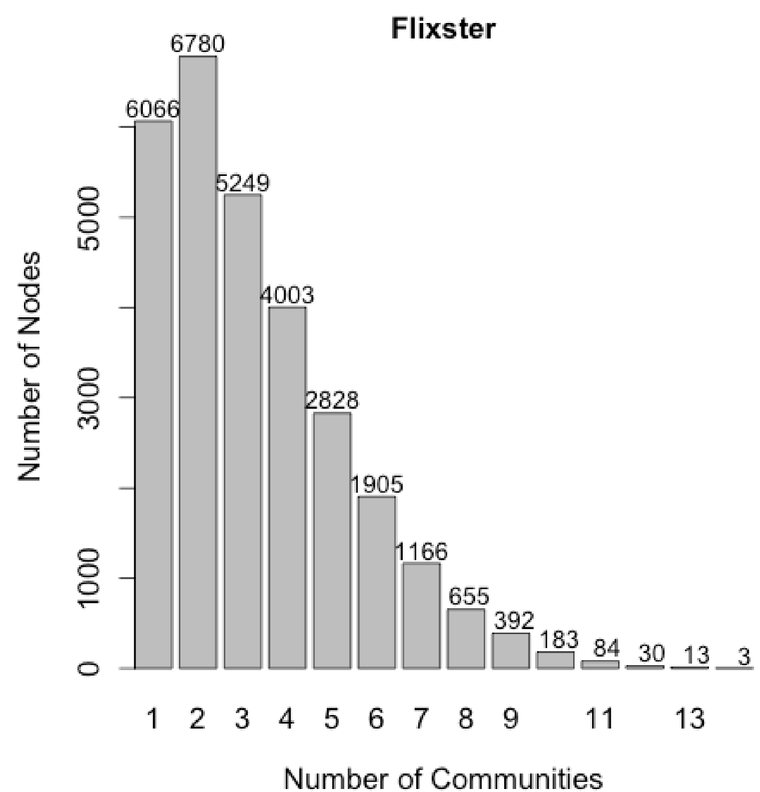
Characterizing the communities

In how many communities users and items tend to participate?

The participation in a community can be inferred by the parameter:

$$\eta_{u,a,k}(\Theta) = P(z_a^k, w_a^u | a \in \mathbb{D}, \Theta)$$



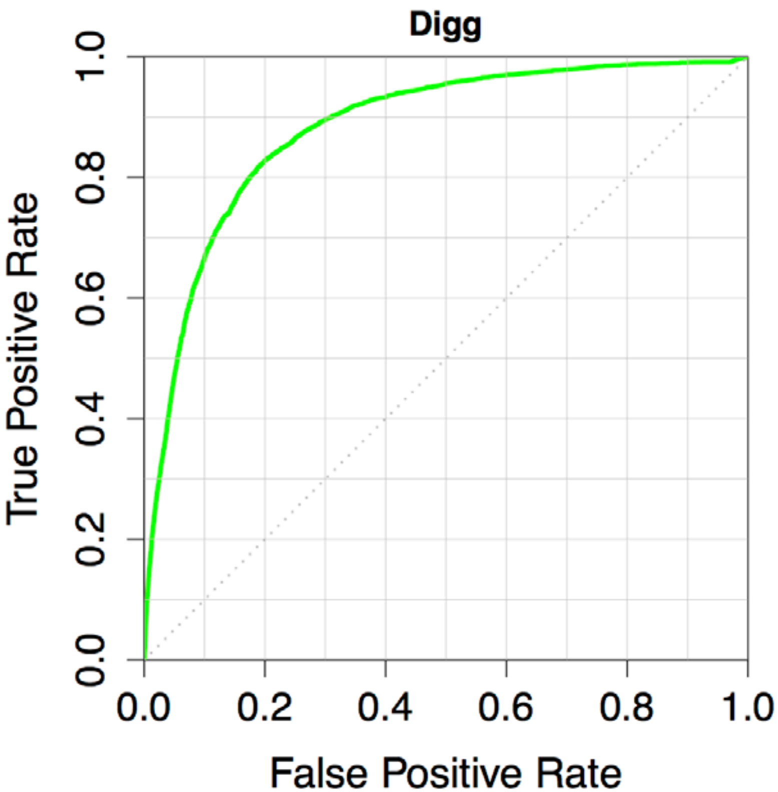


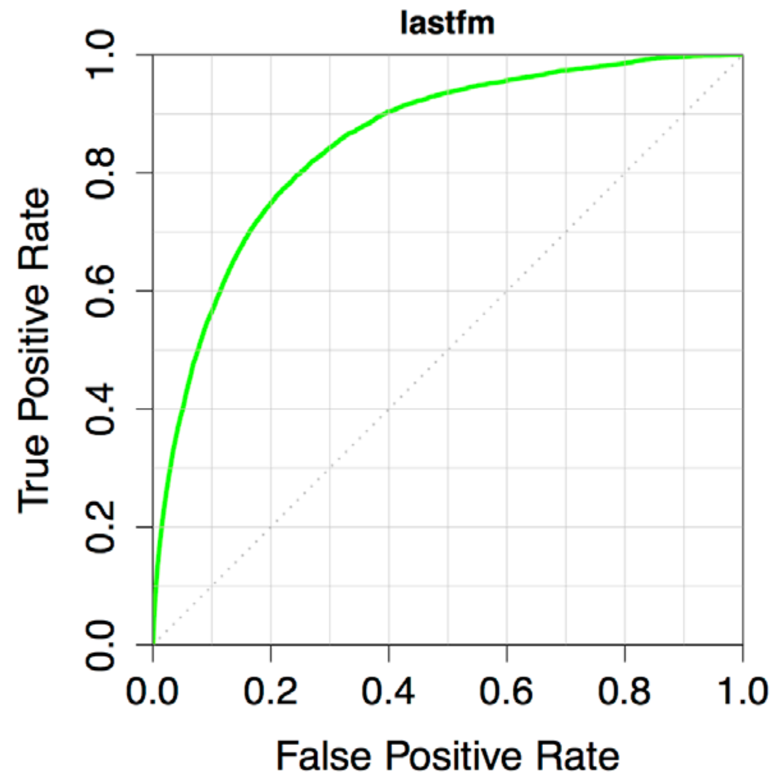
Link Prediction

(Preliminary results to be presented in the extended version)

CCN directly models links probabilities:

$$P(u, v|\Theta) = \sum_k \vartheta_u^k \varphi_v^k \pi_k$$





And what if the social graph is not available?

Detecting communities by mining the propagation log only

"Influence-based Network-oblivious Community Detection"

a.k.a.

"Community detection without the network"

Barbieri, Bonchi, Manco

(ICDM 2013)

Who to Follow and Why: Link Prediction with Explanations

Nicola Barbieri, Francesco Bonchi, Giuseppe Manco

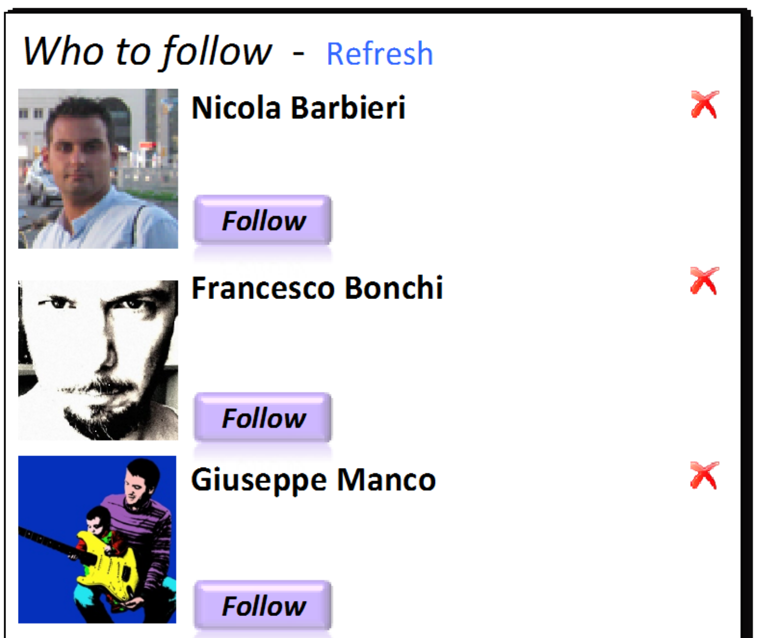
KDD 2014

Motivation

- User recommender systems are a key component in any on-line social networking platform:
  - Assist new users in building their network;
  - Drive engagement and loyalty.

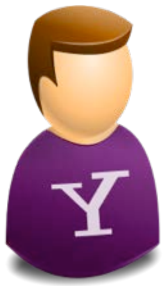
Given a snapshot of a (social) network, can we infer which new interactions among its members are likely to occur in the near future?

Nowell & Kleinberg, 2003



Providing explanations in the context of user recommendation systems is still largely underdeveloped

Modeling socio-topical relationships

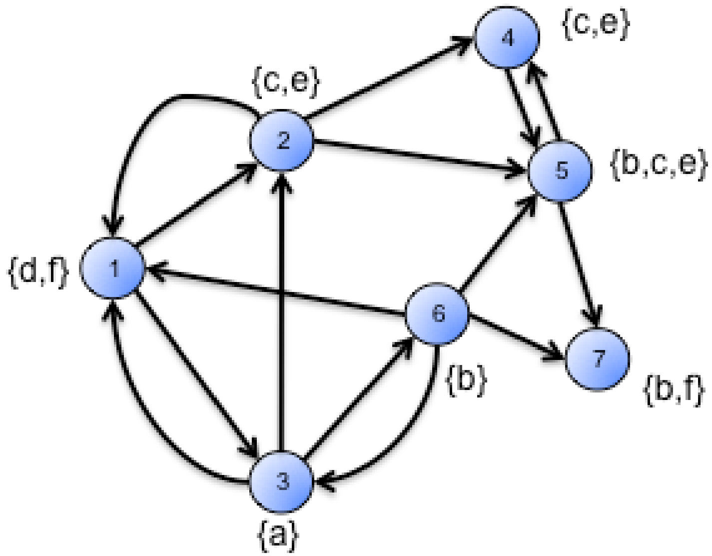


- ✓ Has good friends in Barcelona
- ✓ Does research on web mining
- ✓ Likes blues music

Common identity and common bond theory:

- <u>Identity-based</u> attachment holds when people join a community based on their interest in a well-defined common topic;
- <u>Bond-based</u> attachment is driven by personal social relations with other specific individuals.

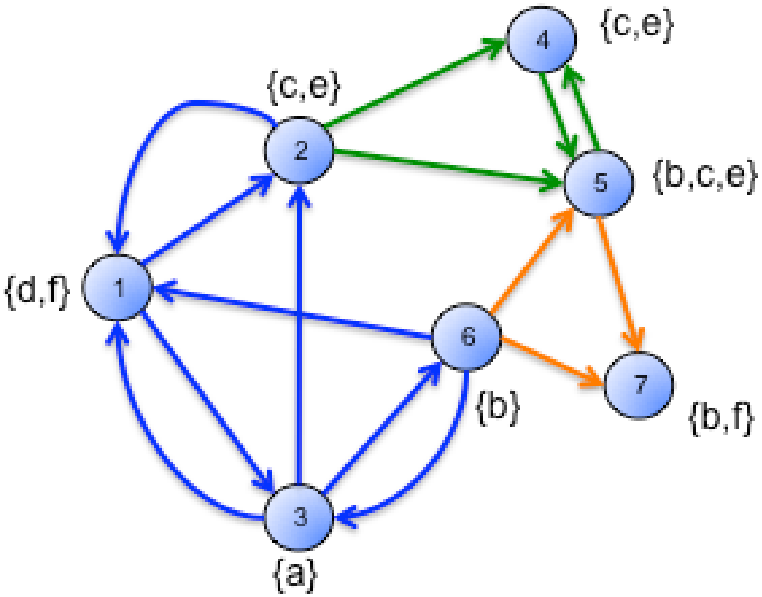
Latent factor modeling of socio-topical relationships



- Directed attributed-graph
- {1,2,3,4,5,6,7} user-set
- Links encode following relationships
- {a,b,c,d,e,f} features adopted by users

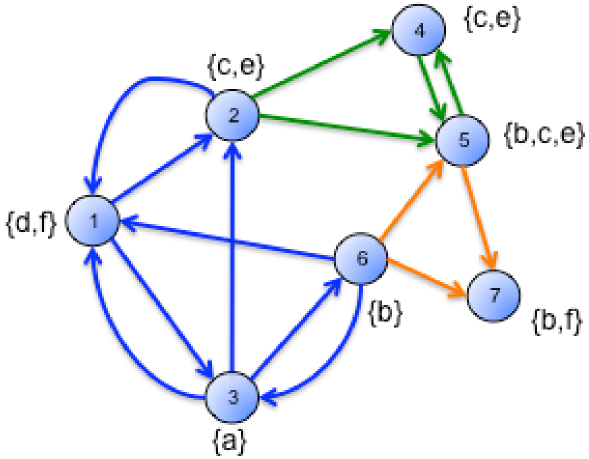
  E.g. hashtags, tags, products purchased

Latent factor modeling of socio-topical relationships



- 3 communities:
  - Blue links are bond-based;
  - Green and orange links are identity-based.
- Bond-based communities tend to have high density and reciprocal links
- Identity-based communities tend to exhibit a clear directionality

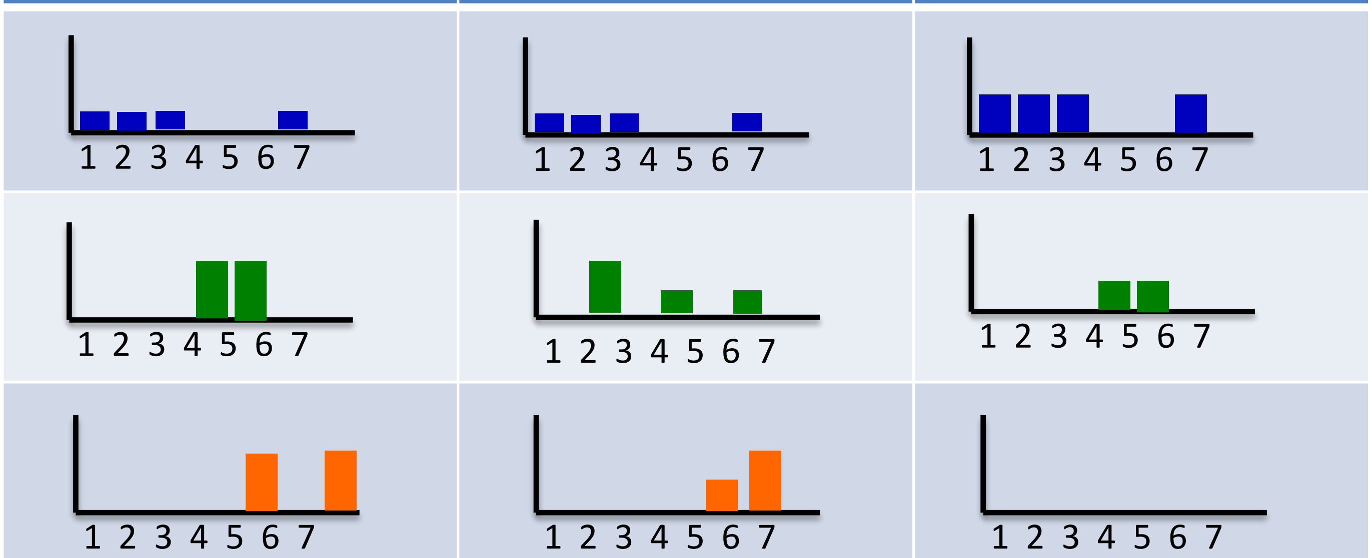
Latent factor modeling of socio-topical relationships



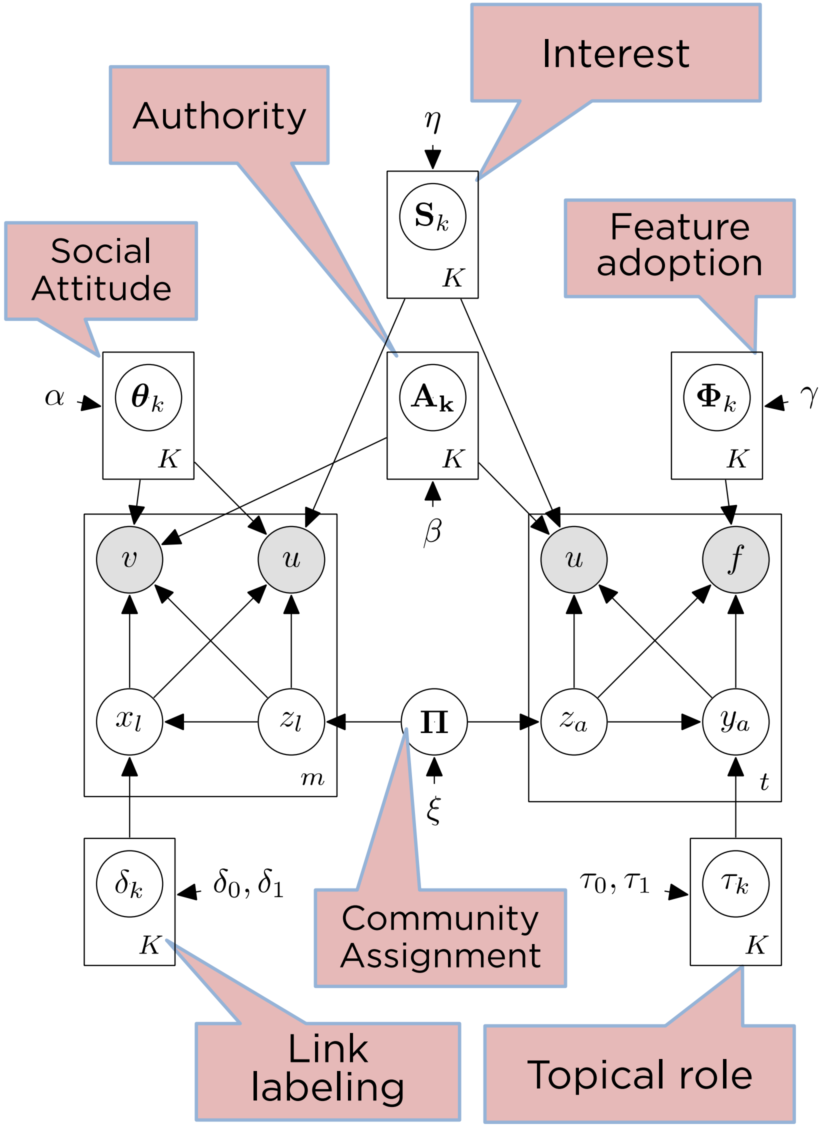
The role and degree of involvement of each user u in the community/topic k is governed by three parameters:

Authority – **Susceptibility (or Interest)** - **Social attitude**

Authority	Susceptibility	Social attitude



WTFW: Generative model



1. sample $\mathbf{\Pi} \sim Dir\left(\vec{\xi}\right)$
2. For each $k \in \{1, \ldots, K\}$ sample

$$\delta_k \sim Beta(\delta_0, \delta_1) \qquad \tau_k \sim Beta(\tau_0, \tau_1)$$
$$\mathbf{\Phi}_k \sim Dir\left(\vec{\gamma}\right) \qquad \boldsymbol{\theta}_k \sim Dir\left(\vec{\alpha}\right)$$
$$\boldsymbol{A}_k \sim Dir\left(\vec{\beta}\right) \qquad \boldsymbol{S}_k \sim Dir\left(\vec{\eta}\right)$$

3. For each link $l \in \{l_1, \ldots, l_m\}$ to generate:
   (a) Choose $k \sim Discrete(\mathbf{\Pi})$
   (b) Sample $x_l \sim Bernoulli(\delta_k)$
   (c) if $x_l = 1$
   - sample source $u \sim Discrete(\boldsymbol{\theta}_k)$
   - sample destination $v \sim Discrete(\boldsymbol{\theta}_k)$

   (d) else
   - sample source $u \sim Discrete(\mathbf{S}_k)$
   - sample destination $v \sim Discrete(\mathbf{A}_k)$
4. For each feature pair $a \in \{a_1, \cdots, a_t\}$ to associate
   (a) sample $k \sim Discrete(\Pi)$
   (b) Sample $y_a \sim Bernoulli(\tau_k)$:
   - if $y_a = 1$ then $u_a \sim Discrete(\boldsymbol{A}_k)$
   - otherwise $u_a \sim Discrete(\boldsymbol{S}_k)$

   (c) sample $f_a \sim Discrete(\mathbf{\Phi}_k)$

Link prediction

- The probability of observing link *l=(u,v)* and the adoption of a feature *a=(u,f)* can be expressed as mixtures over the latent community assignments z_l and z_a:

$$\Pr(l|\Theta) = \sum_{k=1}^{K} \pi_k \Pr(l|z_l = k, \Theta)$$

$$\Pr(l|z_l = k, \Theta) =$$

$$\delta_k \cdot \underbrace{\theta_{k,u} \cdot \theta_{k,v}}_{\text{Social affinity}} + (1 - \delta_k) \cdot \underbrace{S_{k,u} \cdot A_{k,v}}_{\text{Topical affinity}}$$

Takes into account the socio-topical tendency of each community

$$\Pr(a|\Theta) = \sum_{k=1}^{K} \pi_k \Pr(a|z_a = k, \Theta)$$

$$\Pr(a|z_a = k, \Theta) = \underbrace{(\tau_k A_{k,u} + (1 - \tau_k) \cdot S_{k,u})}_{\text{Topical involvement}} \Phi_{k,f}$$

It depends on the degree of topical involvement of the user and by the likelihood of observing the feature within k

Link labeling and explanations

A social link u → v (u should follow v) is recommended when u and v are both members of at least one social community.

$$\Pr(\ (\mathrm{u} \rightarrow \mathrm{v}) \text{ is social}) \propto \sum_{k} \pi_k \cdot \delta_k \cdot \theta_{k,u} \cdot \theta_{k,v}$$

- Explanation can be provided as common friends in the communities that better explain the link.

A topical link **u** → **v** is recommended to (u) when (v) is authoritative in a topic on which (u) has shown interest.

$$\Pr(\ (\mathrm{u} \rightarrow \mathrm{v}) \text{ is topical}) \propto \sum_{k} \pi_k \cdot (1 - \delta_k) \cdot S_{k,u} \cdot A_{k,v}$$

- Explanation as a list of features that characterize the authoritativeness of (v) in (u)'s topics of interest.

Evaluation

- On both Twitter and Flickr the link creation process can be explained in terms of interest identity and/or personal social relations.
- Features:
  - On Twitter: all hashtags and mentions adopted by the user;
  - On Flickr: all the tags assigned by the user.

	Twitter	Flickr
Number of nodes	81,306	80,000
Number of links	1,768,149	14,036,407
Number of one-way links	1,342,311	9,604,945
Number of bidirectional links	425,838	4,431,462
Number of social links	-	6,747,085
Number of topical links	-	7,289,322
Avg in-degree	21	175
Avg out-degree	25	181
Number of features	211,225	819,201
Number of feature assignments	1,102,000	37,316,862
Avg. features per user	15	613
Avg. users per feature	5	45

- Flickr contains ground-truth for the labeling relationships.
- Relationships flagged as either “family” or “friends” are labeled as social, the remaining ones as topical.

Accuracy on link prediction

- Evaluation setting:
  - On Twitter: Monte Carlo 5 Cross-Validation;
  - On Flickr: Chronological split.
- Negative samples: all the 2-hops non-existing links.
- Competitors:
  - Common neighbors and features;
  - Adamic-Adar on neighbors and features;
  - Joint SVD on the combined adjacency/feature matrices

Accuracy on link prediction

Method	Split	8	16	32	64	128	256
		Number of latent factors					
WTFW	60/40	0.567	0.615	0.667	0.707	0.739	**0.792**
	70/30	0.565	0.631	0.680	0.713	0.749	**0.798**
	80/20	0.586	0.639	0.692	0.732	0.760	**0.812**
JSVD	60/40	0.439	0.471	0.525	0.588	0.660	0.768
	70/30	0.446	0.48	0.537	0.602	0.679	0.744
	80/20	0.454	0.495	0.545	0.617	0.693	0.763
CNF		0.7025/0.7125/0.7199					
AA-NF		0.7301/0.7397/0.7472					

Table 3: AUC on link prediction - `Twitter`

Method	8	16	32	64	128	256
	Number of latent factors					
WTFW	0.6467	0.6488	0.6534	0.6576	0.661	**0.677**
JSVD	0.598	0.596	0.597	0.609	0.619	0.624
CNF	0.53					
AA-NF	0.58					

Table 4: AUC on link prediction - `Flickr`

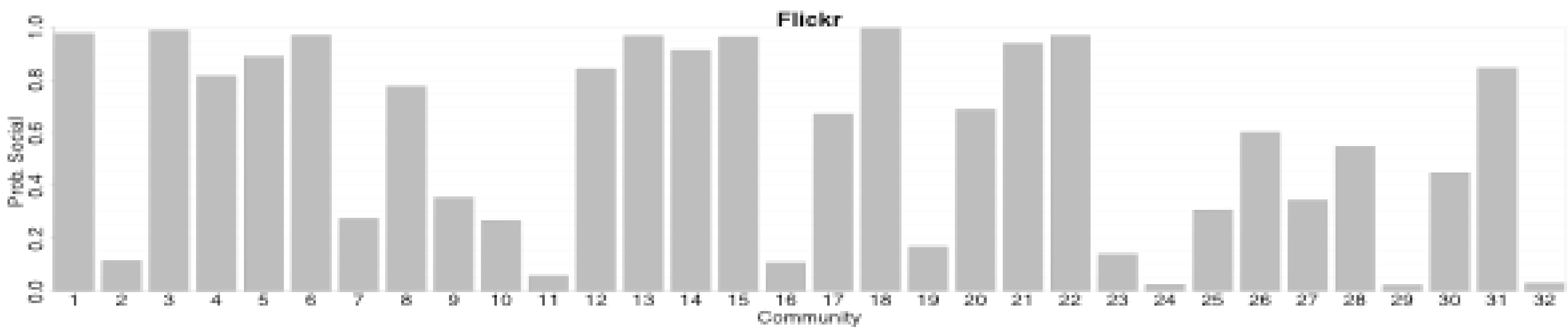
Link labeling

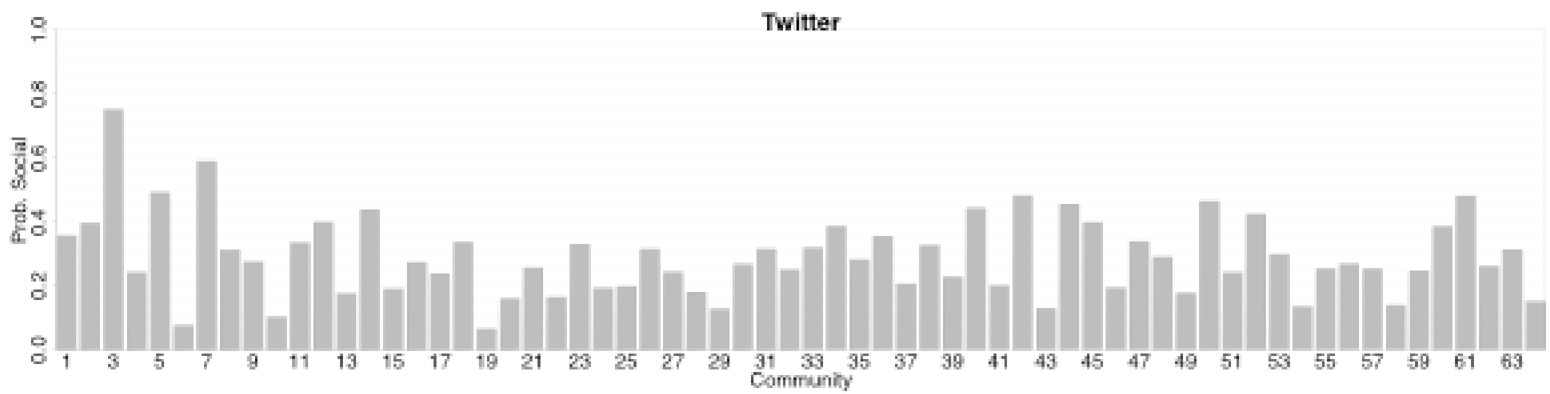
- Baseline on Link Labeling

$$\Pr(x_l = 1|l) = \frac{|N(u) \cap N(v)|}{|N(u) \cap N(v)| + |F(u) \cap F(v)|}$$

Method	Number of latent factors: 8	16	32	64	128	256
WTFW	0.7393	0.7548	**0.7603**	0.6883	0.6618	0.6582
Baseline	0.6545					

AUC on link labeling - `Flickr`.





Anecdotal evidence

Feature	Prob. Social
birthday	0.69
family	0.67
wedding	0.69
party	0.67
puppy	0.69

Feature	Prob. Social
hdr	0.40
vintage	0.29
collage	0.24
nude	0.08
polaroid	0.28

Table 6: Social/Topical connotations of selected tags on `Flickr`.

Flickr			
Topic 1 $\delta = 0.98$	*Topic 5* $\delta = 0.98$	*Topic 18* $\delta = 0.17$	*Topic 22* $\delta = 0.14$
Christmas, esther, passenger, Birthday, eros, party, stories, apple, curling, homemade	family, mom, dog, driving, vitus, bakery, woods, birthday, friends, halloween, shirt, brothers, baby	handmade, warehouse, vintage, knitting, craft, green, pansies, doll, sewing	bird, art, design, illustration, drawing, fo-toincatenate, sketch, street, painting, ink, graffiti

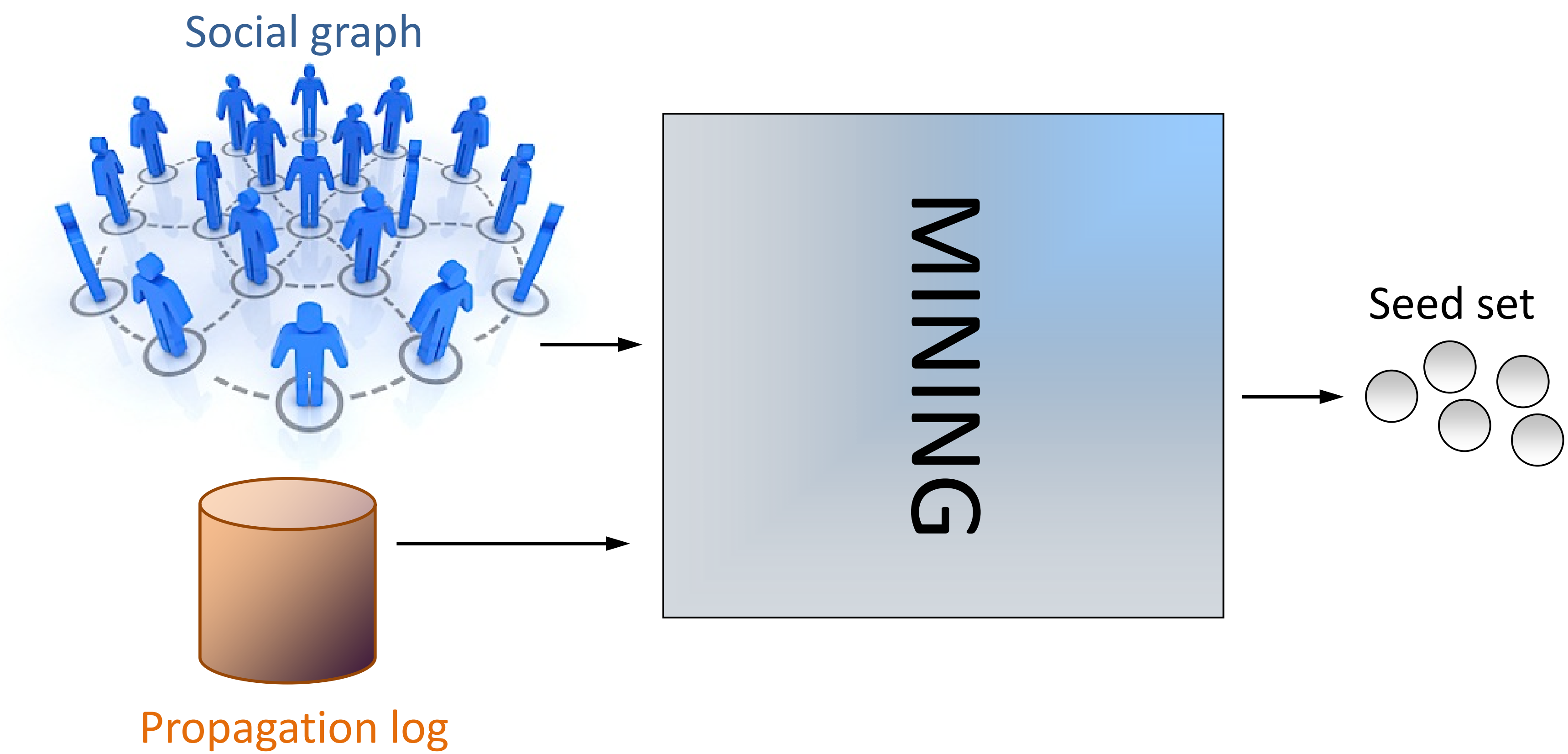
Twitter			
Topic 3 $\delta = 0.74$	*Topic 9* $\delta = 0.27$	*Topic 64* $\delta = 0.16$	*Topic 47* $\delta = 0.33$
TeamFollow-Back TFB FollowNGain fb InstantFol-lowBack nowplaying lastfm Tea-mAutoFollow Follow4Follow 500aDay anime 4sqDay	Autodesk BIM AutoCAD Revit AU2012 Civil3D AEC adsk_sf2012 SWTOR revit CAD au2011 cloud 3dsMax AU2011 C3D2013	ISS space science Discovery Mars nasa spottheshut-tle ESA astronomy Enterprise Soyuz	Game-ofThrones FakeWesteros GoT ooc SXSWesteros TheGhostofHar-renhal Gardenof-Bones asoiaf GRRM GOT

Table 7: Most representative features of selected communities.

Thank you!
Questions?

@FrancescoBonchi
www.francescobonchi.com
francescobonchi@acm.org

Another approach: direct mining!



Influential users: direct mining methods

A. Goyal, F. Bonchi, L. V. S. Lakshmanan

Discovering leaders from community actions (CIKM 2008)

A. Goyal, B. W. On, F. Bonchi, L. V. S. Lakshmanan

GuruMine: a Pattern Mining System for Discovering Leaders and Tribes (ICDE 2009)

A. Goyal, F. Bonchi, L. V. S. Lakshmanan

A Data-Based Approach to Social Influence Maximization (VLDB 2012)

Sparsification of Influence Networks

which connections are most important
for the propagation of actions?

keep only important connections

data reduction
visualization
clustering
efficient graph analysis
find the backbone of influence/information networks

Influence-driven sparsification

M. Mathioudakis, F. Bonchi, C.Castillo, A. Gionis, A. Ukkonen
Sparsification of Influence Networks (KDD 2011)

F. Bonchi, G. De Francisci Morales, A. Gionis, A. Ukkonen
Activity Preserving Graph Simplification (DAMI journal 2013)

Sparsification

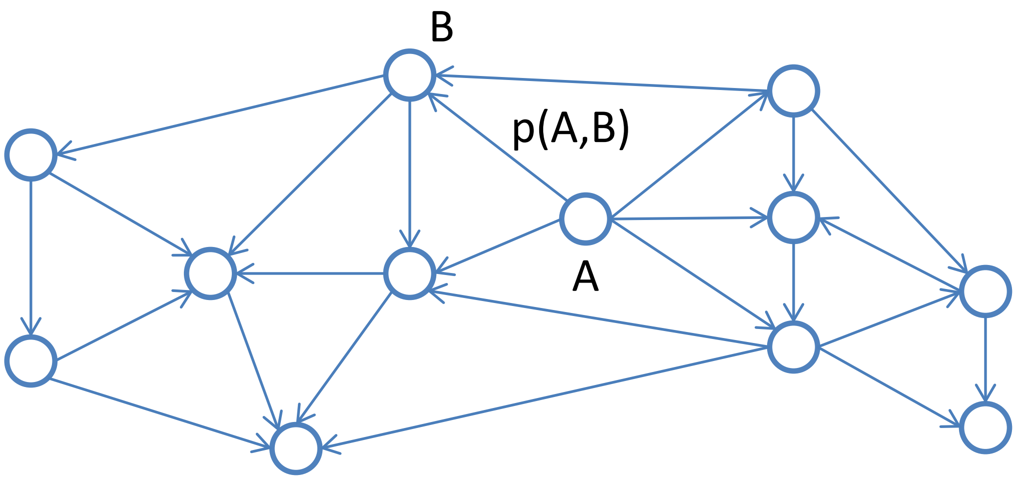
social network

set of propagations

p(A,B)

k arcs

most likely to
explain propagations
(assuming the Independent Cascade model)



Sparsification

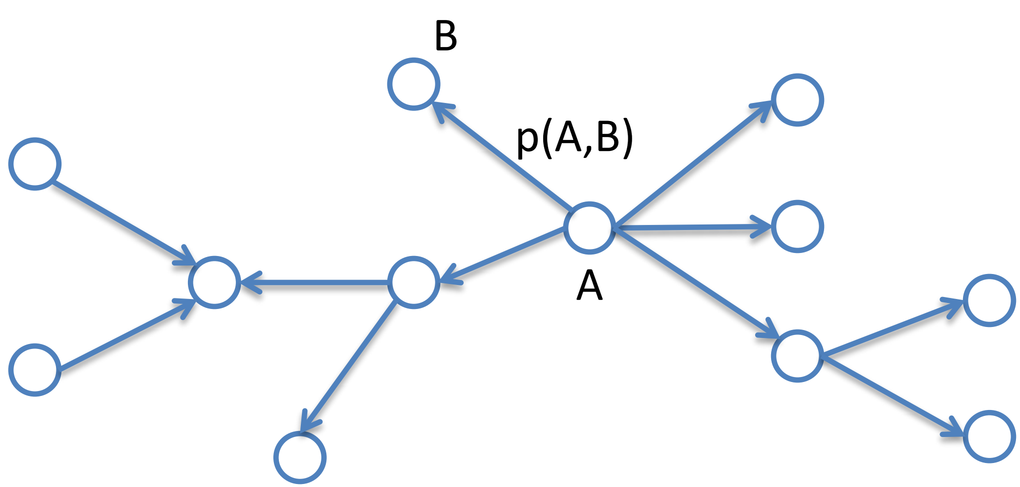
social network

set of propagations

p(A,B)

k arcs

most likely to explain propagations (assuming the Independent Cascade model)



Solution

not the k arcs with largest probabilities!

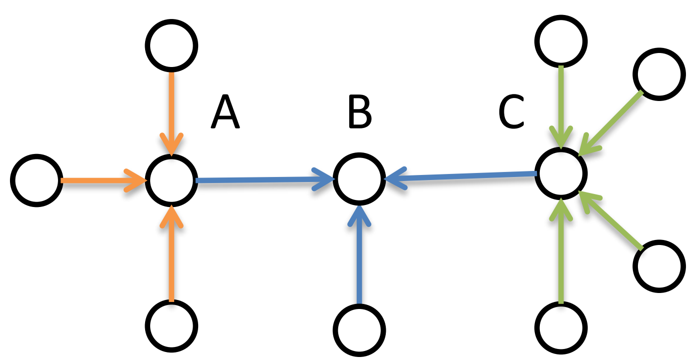
problem is NP-hard and inapproximable

sparsify separately incoming arcs of individual nodes

optimize corresponding likelihood

dynamic programming

optimal solution



$$k_A + k_B + k_C = k$$

Spine - **sp**arsification of **i**nfluence **ne**tworks

http://www.cs.toronto.edu/~mathiou/spine/

greedy algorithm

two phases

phase 1

obtain a non-zero-likelihood solution

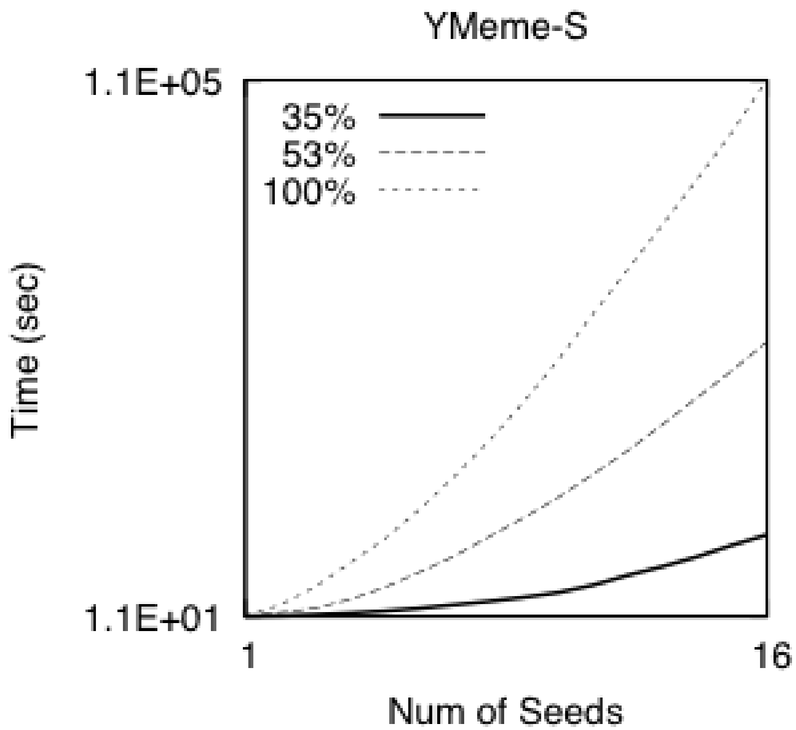
(greedy algorithm for Hitting Set problem)

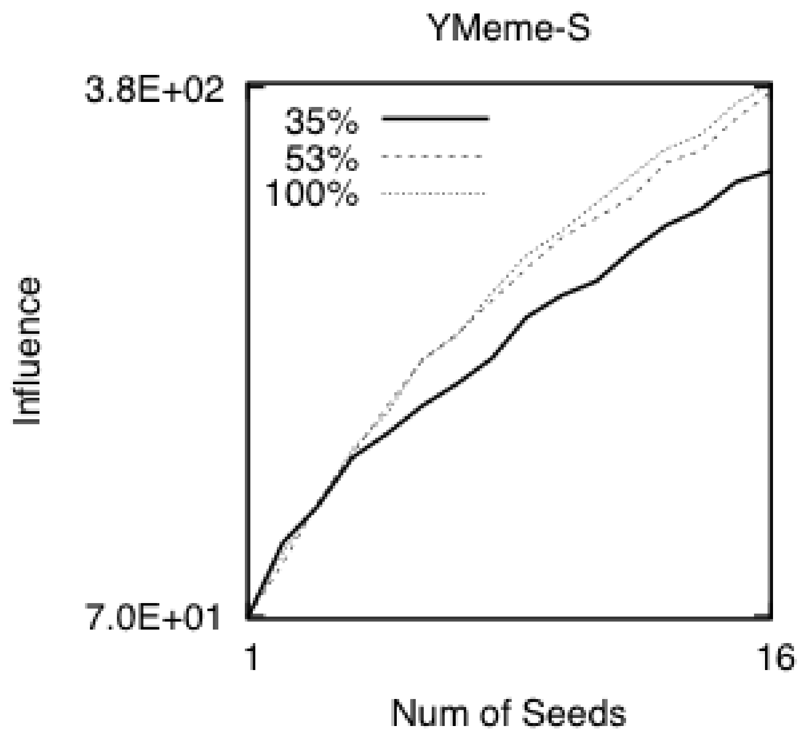
phase 2

add one arc at a time, the one that offers largest increase in likelihood

(approximation guarantee for phase 2 thanks to submodularity)

Application to Influence Maximization





Same setting, other objectives

A. Goyal, F. Bonchi, L. Lakshmanan, S. Venkatasubramanian (SNAM journal)
On Minimizing Budget and Time in Influence Propagation over Social Networks

F. Bonchi, C.Castillo, D. Ienco
The Meme Ranking Problem: Maximizing Microblogging Virality
(ICDM 2010 workshop + Journal of Intelligent Information Systems)

I. Mele, F. Bonchi, A. Gionis (CIKM 2012)
The early-adopter graph and its application to web-page recommendation

W. Lu, F. Bonchi, A. Goyal, L. V. S. Lakshmanan (KDD 2013)
The Bang for the Buck: Fair Competitive Viral Marketing from the Host Perspective

N. Barbieri, F. Bonchi
Influence Maximization with Viral Product Design (SDM 2014)

Summaries and indexes

L. Macchia, F. Bonchi, F. Gullo, L. Chiarandini

Mining Summaries of Propagations (ICDM 2013)

A. Khan, F. Bonchi, A. Gionis, F. Gullo

Fast Reliability Search in Uncertain Graphs (EDBT 2014)

C. Aslay, N. Barbieri, F. Bonchi, R. Baeza-Yates

Online Topic-aware Influence Maximization Queries (EDBT 2014)

Position paper

F. Bonchi

Influence Propagation in Social Networks: A Data Mining Perspective

(IEEE Intelligent Informatics Bulletin)