On information propagation, social influence, and communities

@FrancescoBonchi
www.francescobonchi.com
francesco.bonchi@isi.it

INSTITUTE
FOR SCIENTIFIC INTERCHANGE
FOUNDATION

New Affiliations

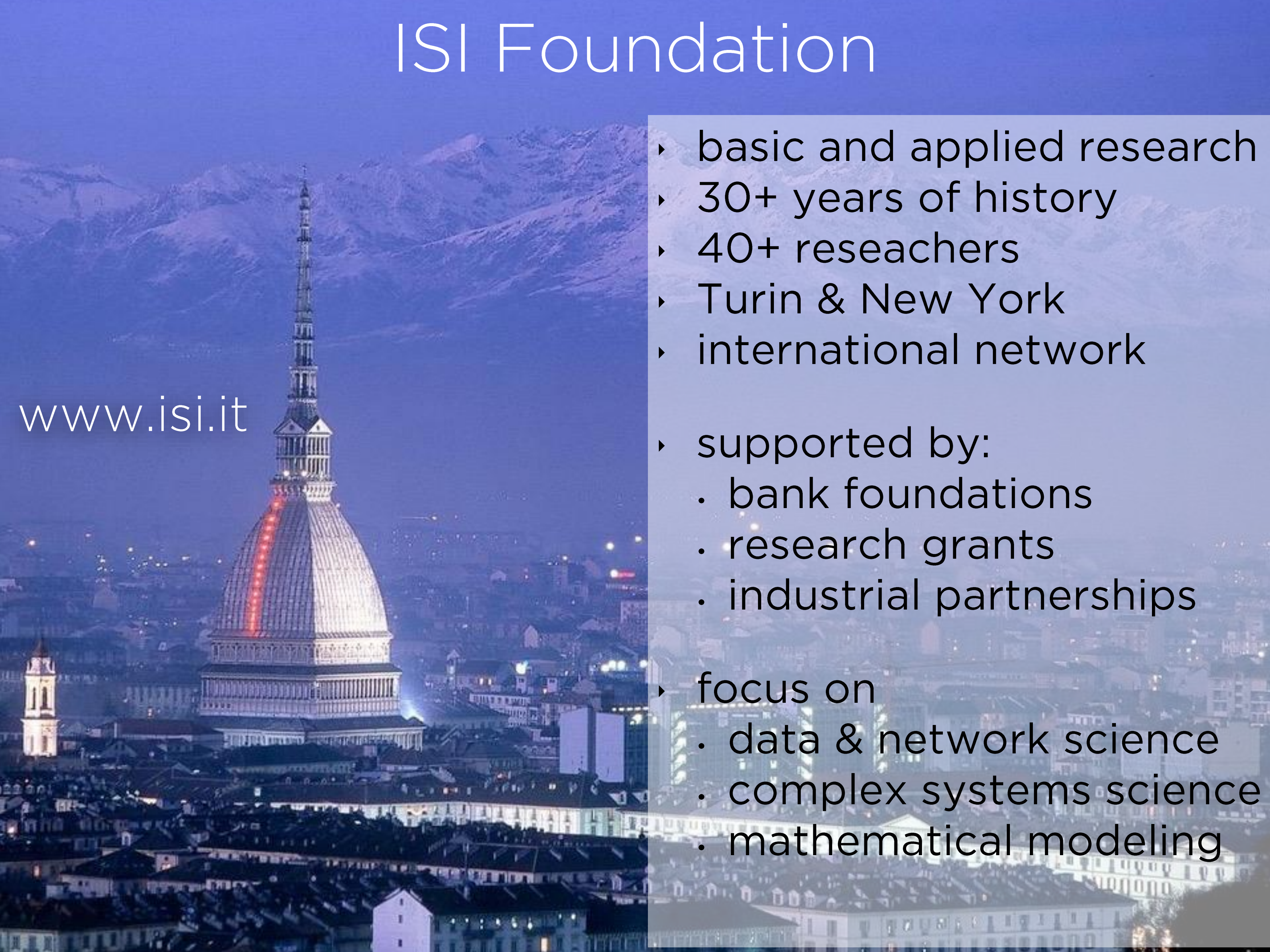
- Algorithmic Data Analytics team @ ISI Foundation (Institute for Scientific Interchange) Turin, Italy http://www.isi.it/

ISI Foundation

- Data Science team @ Eurecat (Technological Center of Catalonia) Barcelona, Spain http://eurecat.org/

eurecat
Centre Tecnològic de Catalunya

ISI Foundation



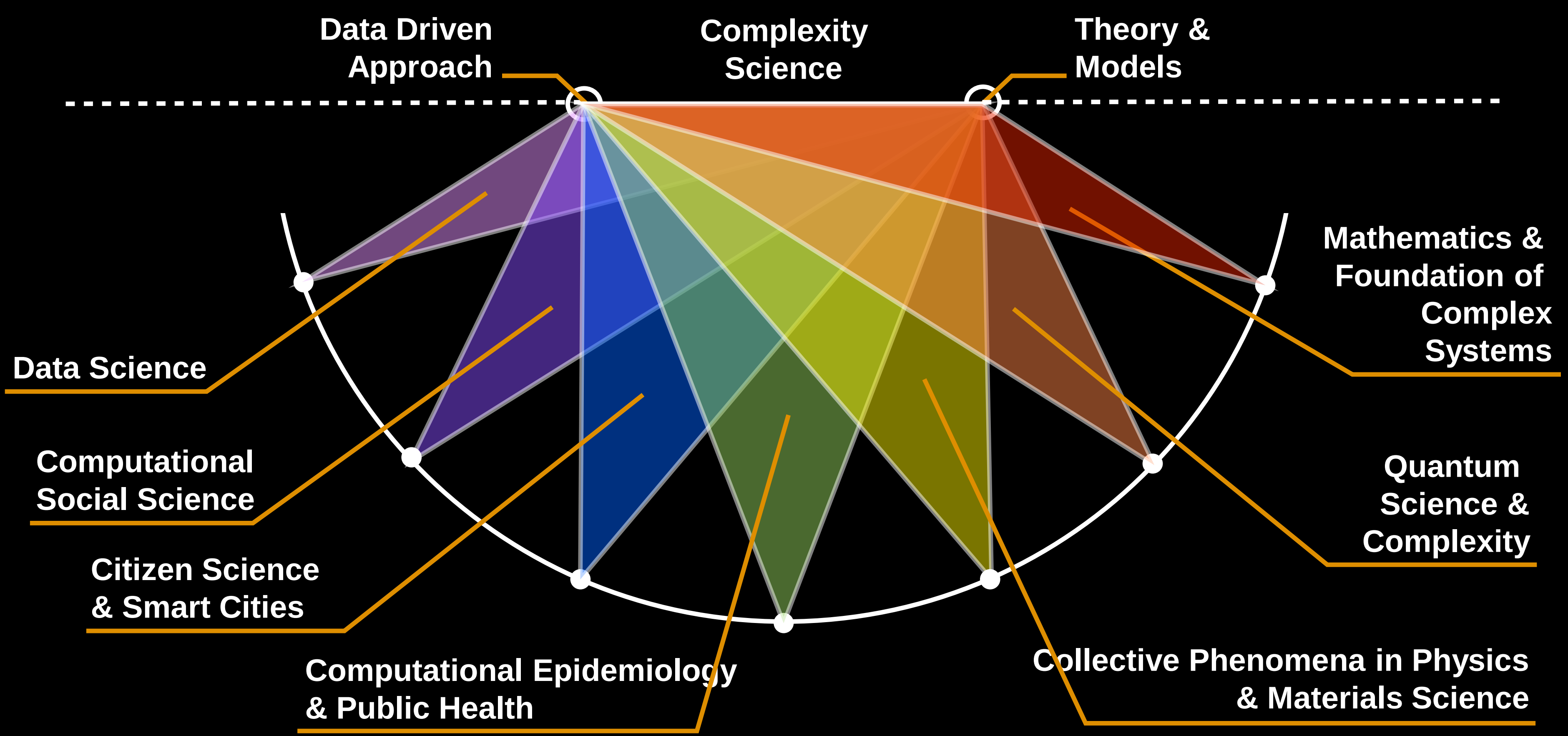
www.isi.it

- basic and applied research
- 30+ years of history
- 40+ reseachers
- Turin & New York
- international network

- supported by:
  - bank foundations
  - research grants
  - industrial partnerships

- focus on
  - data & network science
  - complex systems science
  - mathematical modeling

ISI Foundation:
The Arc of Science



BARCELONA

ICDM 2016

IEEE International Conference on Data Mining

Paper submission: June 17

Conference: December 12-15

Plan of the talk

- Background: information propagation in social networks, influence maximization, learning the strenght of social influence

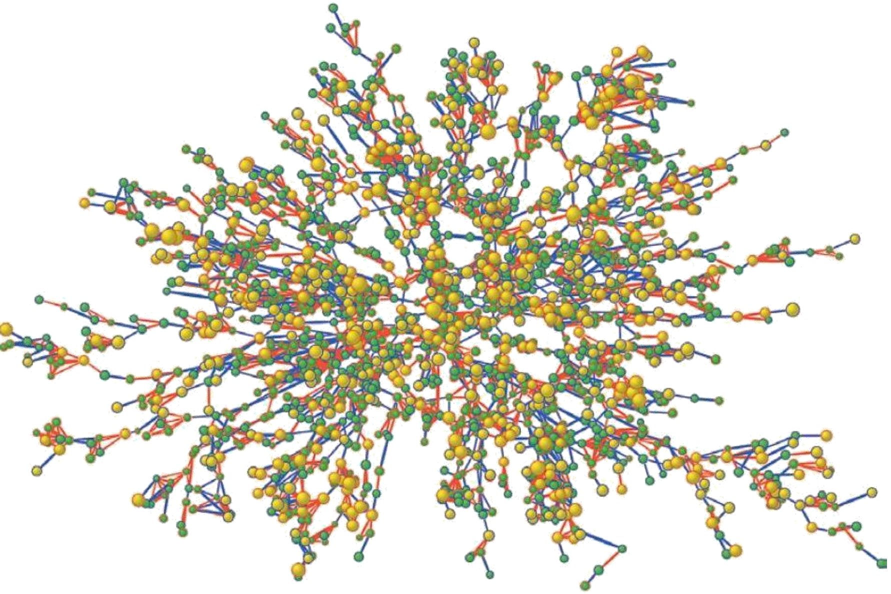
- Part I: Adding topic-awareness
- Part II: Cascades & Communities

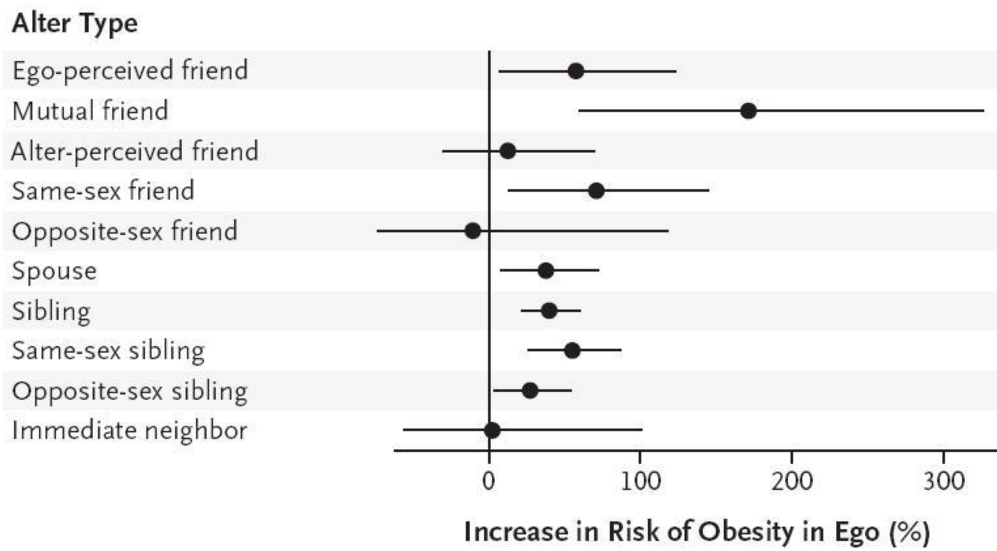
- Concluding remarks

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

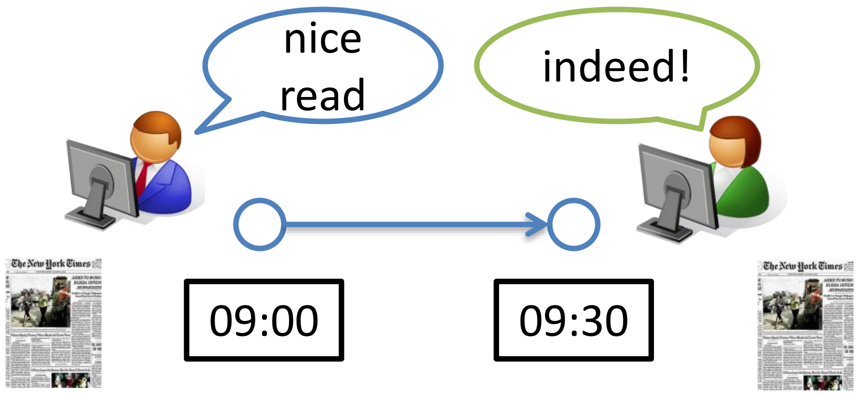
Influence is a causal process

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 from NYU].

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

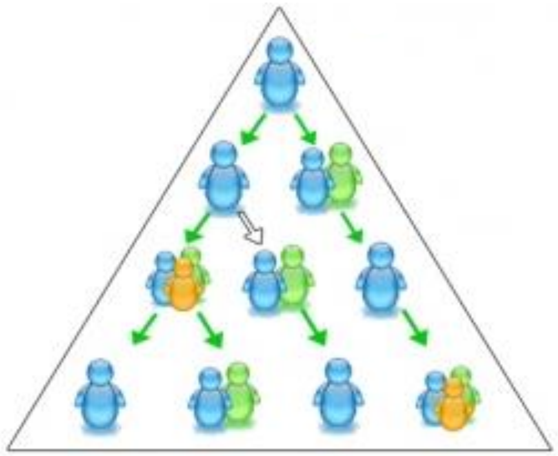
behavioral targeting

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 15 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"*
[more than 3500 citations]

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

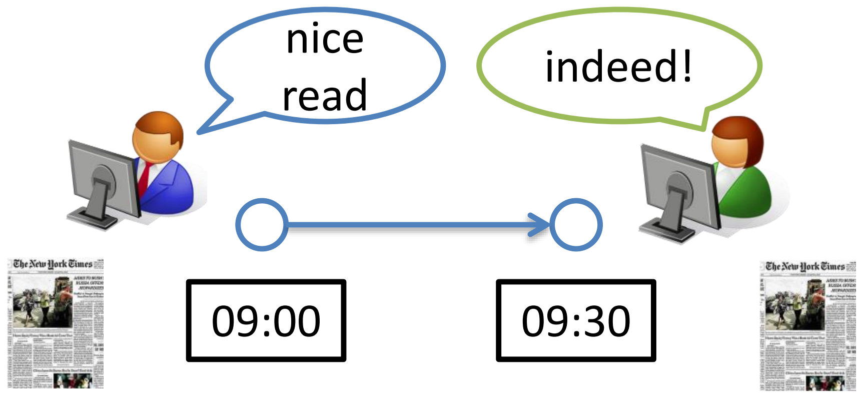
Problem: Given social network *G* with arcs probabilities, budget k, find k-node set *S* that maximizes $\sigma_M(S)$

Two major propagation models considered:

independent cascade (IC) model

linear threshold (LT) model

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

Influence-driven information propagation in on-line social networks

Saito, Nakano, and Kimura (KES'08)
"Prediction of information diffusion probabilities for independent cascade model"

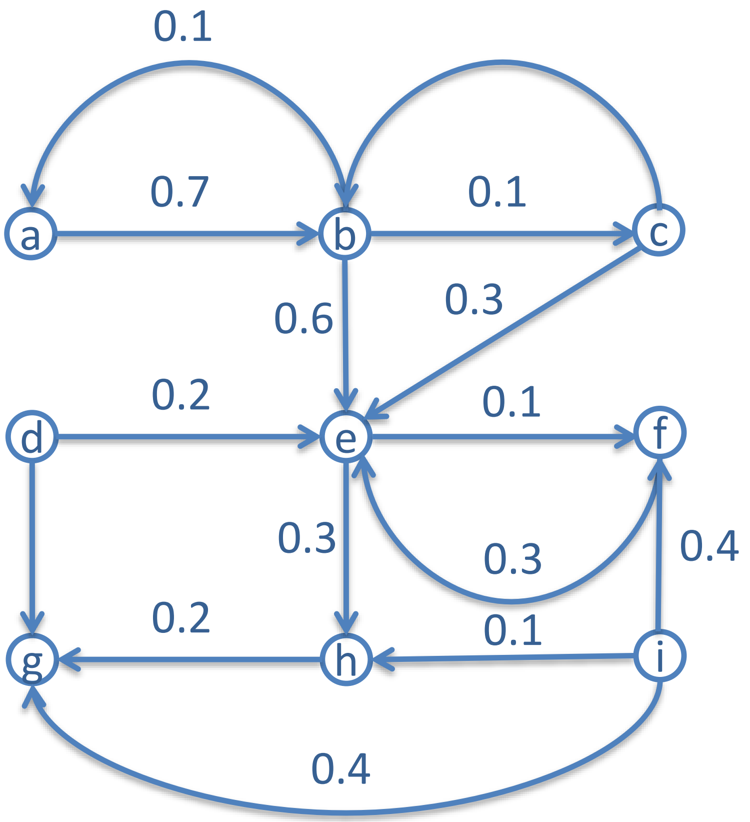
Goyal, Bonchi, and Lakshmanan (WSDM'10)
"Learning influence probabilities in social networks"

Kutzkov, Bifet, Bonchi, and Gionis (KDD'13)
"STRIP: Stream Learning of Influence Probabilities"

Tassa and Bonchi (EDBT'14)
"Privacy Preserving Estimation of Social Influence"

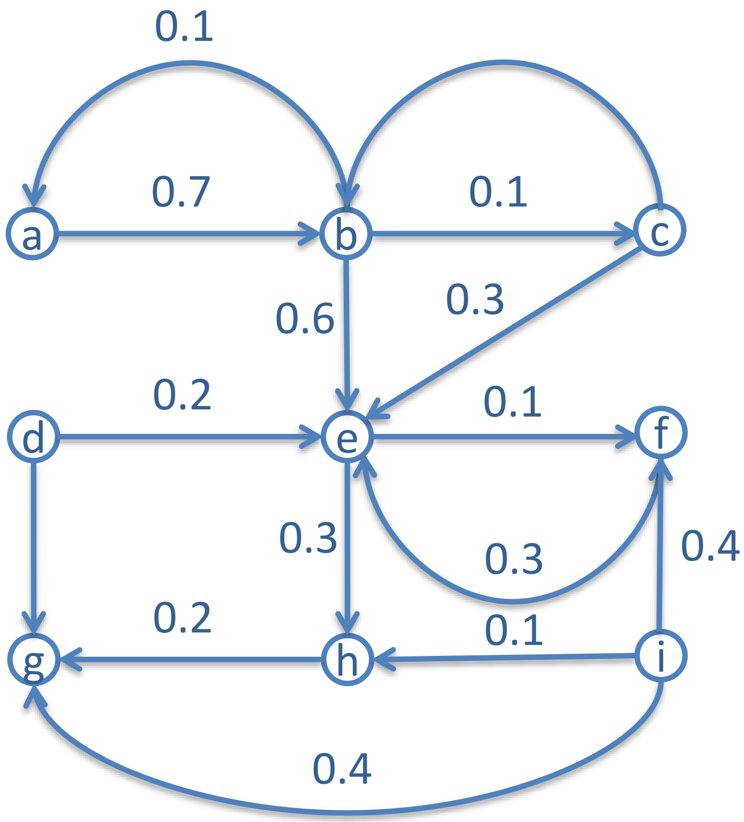

0.7

Independent Cascade model (IC)



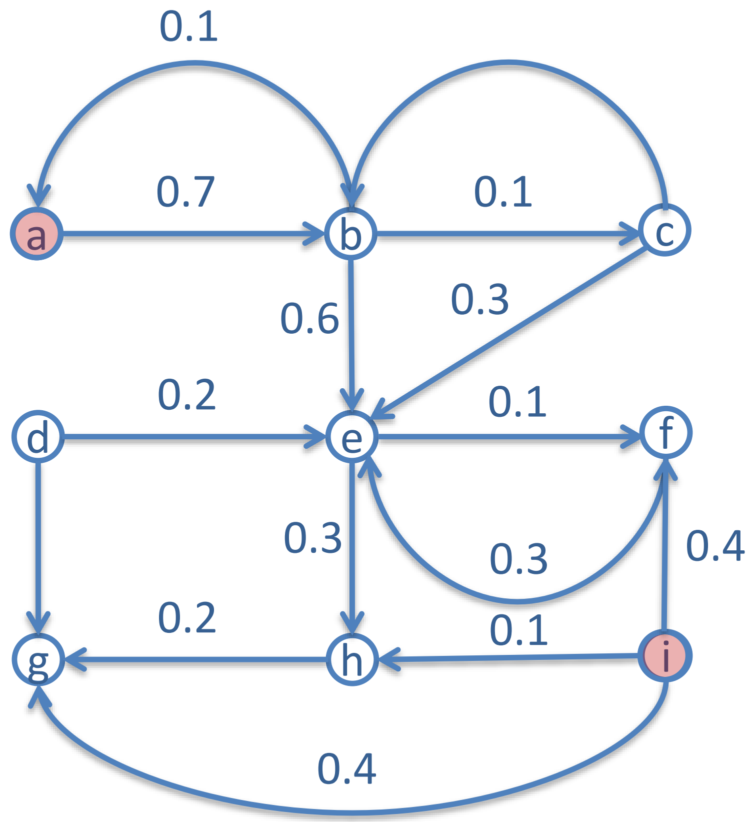
Independent Cascade model (IC)

- 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 the probability on the arc



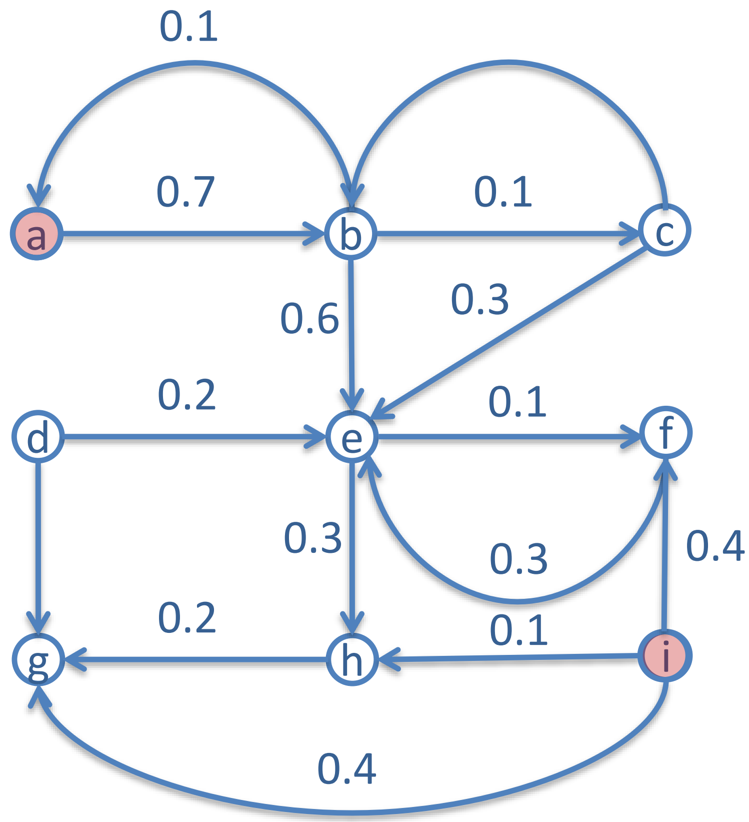
Independent Cascade model (IC)

- 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 the probability on the arc



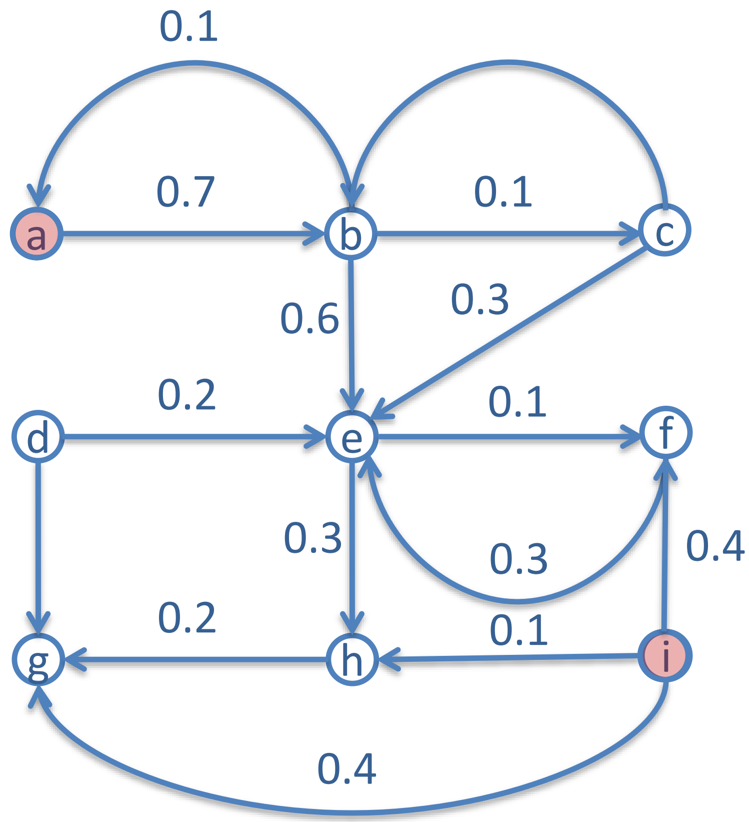
Independent Cascade model (IC)

- 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 the probability on the arc



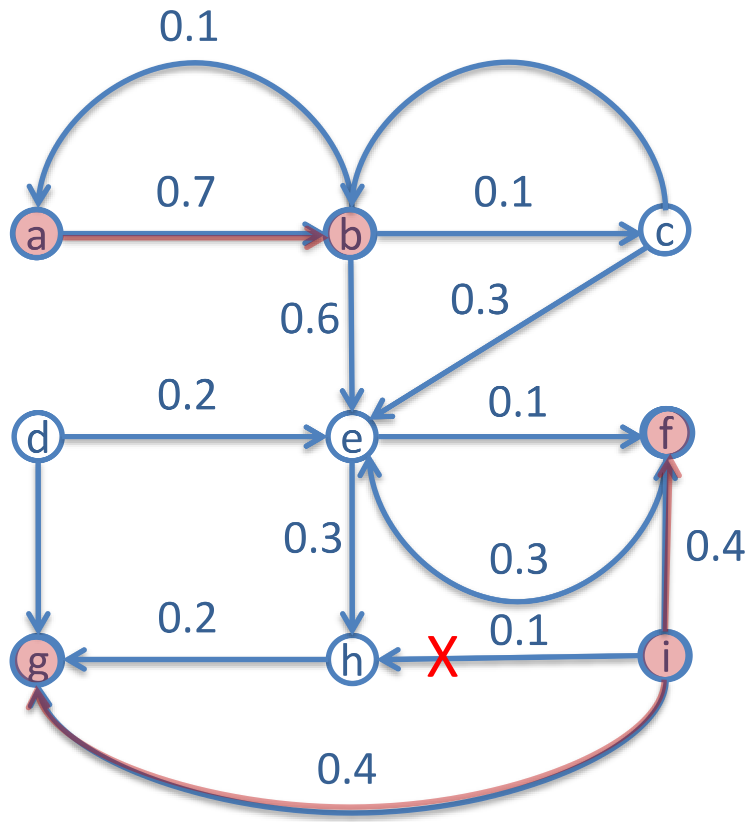
Independent Cascade model (IC)

- 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 the probability on the arc



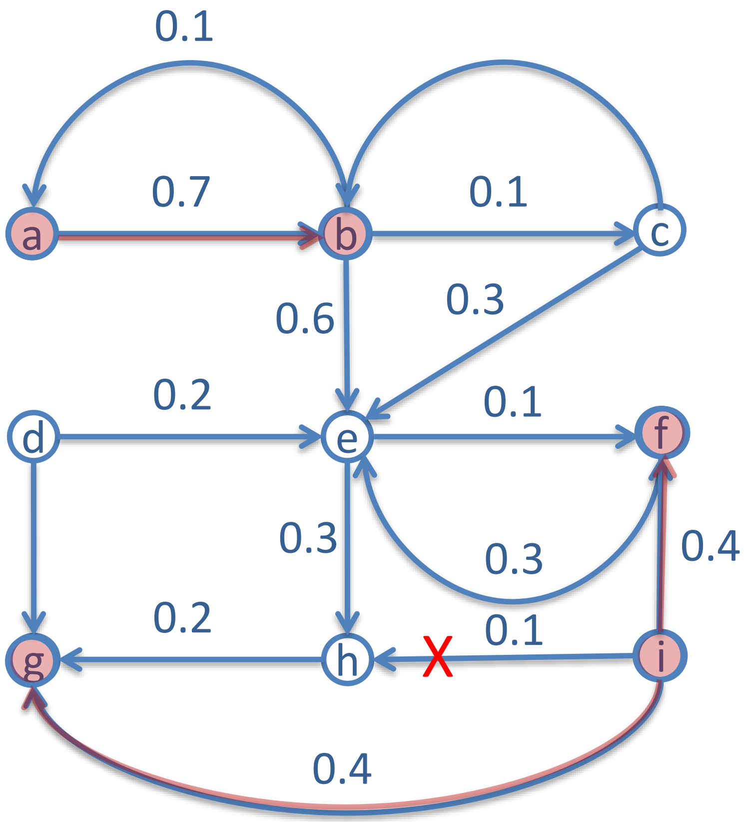
Independent Cascade model (IC)

- 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 the probability on the arc



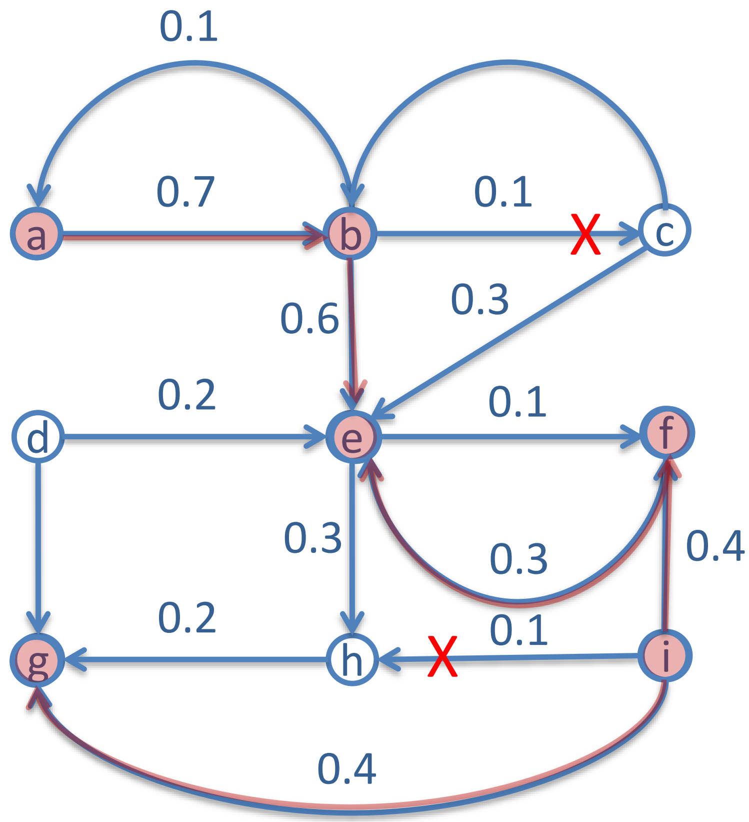
Independent Cascade model (IC)

- 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 the probability on the arc



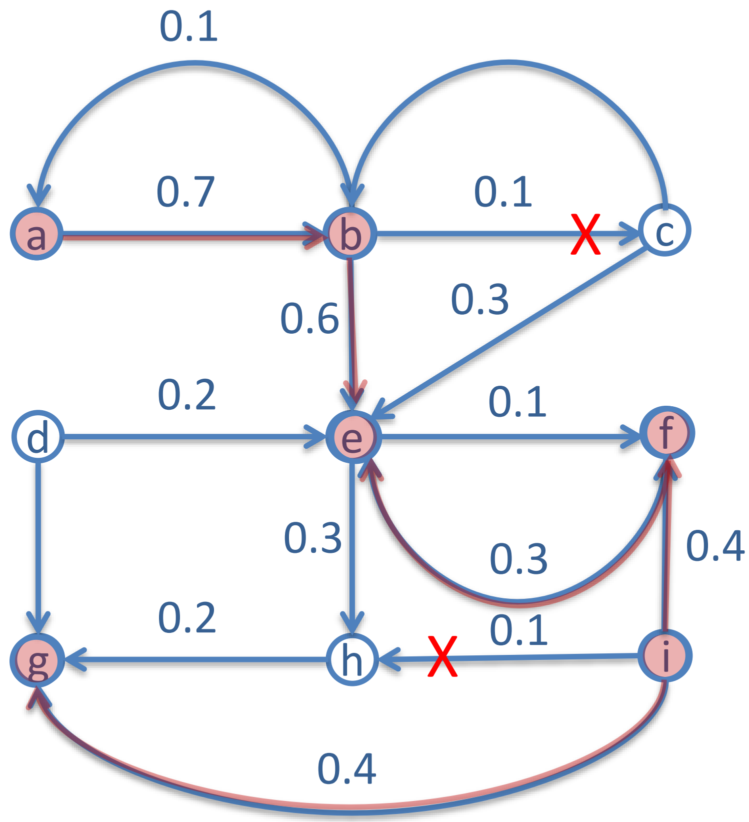
Independent Cascade model (IC)

- 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 the probability on the arc



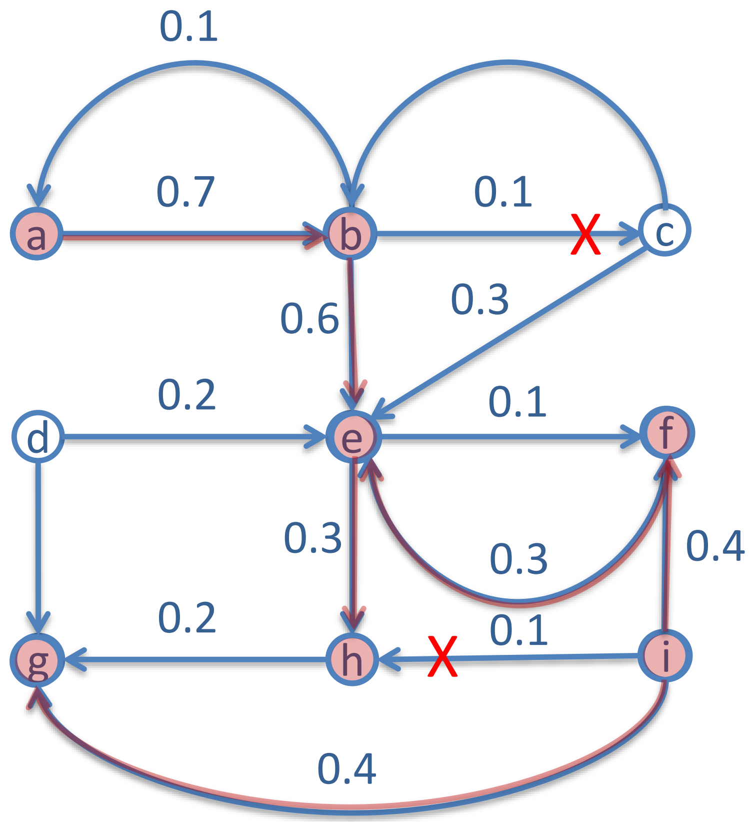
Independent Cascade model (IC)

- 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 the probability on the arc



Independent Cascade model (IC)

- 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 the probability on the arc



Known Results

Bad news: NP-hard optimization problem

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

Theorem: The (1- 1/e) approximation ratio cannot be further improved

Bad news: computing $\sigma_M(S)$ is #P-hard under the IC model

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)

Part I:
Adding topic-awareness

N. Barbieri, F. Bonchi, G. Manco
“Topic-aware Social Influence Propagation Models” (ICDM 2012) (KAIS)

C. Aslay, N. Barbieri, F. Bonchi, R. Baeza-Yates
“Online Topic-aware Influence Maximization Queries” (EDBT 2014)

Topic-aware Social Influence Propagation Models

Barbieri, Bonchi, Manco (ICDM'12 + KAIS)

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

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)

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

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

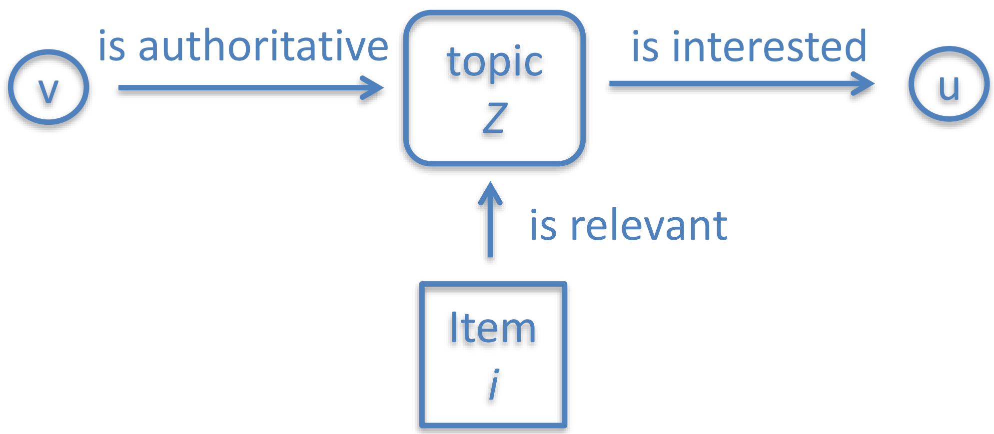
... but:
TIC has a huge number of parameters
#topics(#links + #items)

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

Idea: use topics as a proxy for defining social influence

$$p^i_{v,u}$$



number of parameters:
#topics(2 #nodes + #items)

[Model definition and learning: see details in the paper (!)]

Topic-aware Influence Maximization (TIM) Queries

C. Aslay, N. Barbieri, F. Bonchi, R. Baeza-Yates (EDBT 2014)

- Given
  - a social graph $G = (V,E)$
  - a space of Z topics
  - topic-specific peer-influence probabilities on arcs, $p^{z}_{u,v}$
  - a query item q, $\vec{\gamma}_q$
  - budget k
  - And assuming TIC propagation model

- TIM query asks to find a seed set S of k nodes that maximizes the expected number of nodes adopting item q in the network:

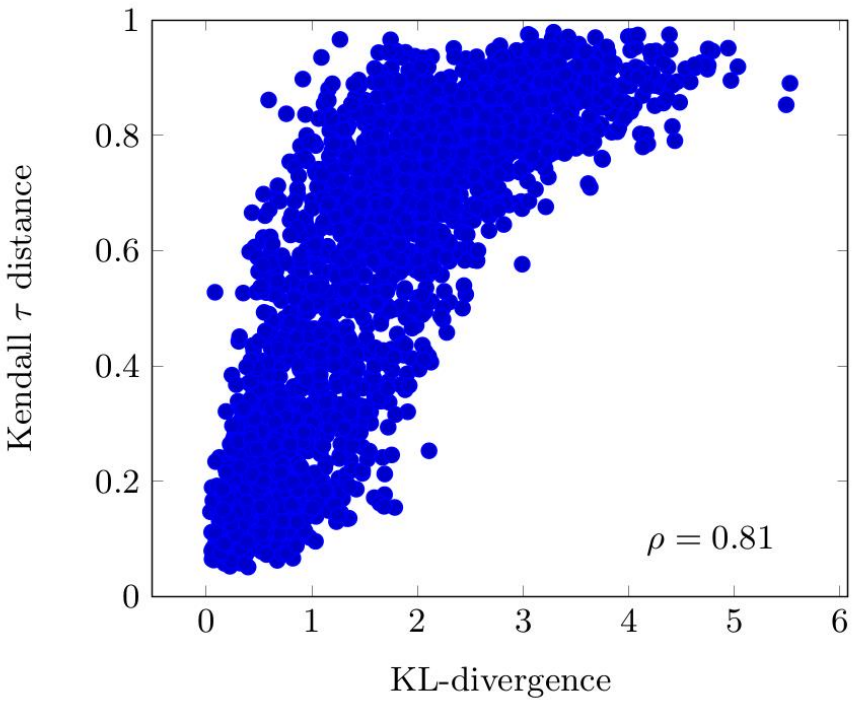
$$Q(\vec{\gamma}_q, k) = \arg\max_{S \subseteq V, |S|=k} \sigma(S, \vec{\gamma}_q)$$

On-line TIM Queries

- TIM query can be processed by standard influence maximization algorithms:
  - Reduce TIC to IC via the derived graph $G^q = (V,A,p^q)$
  - Enjoy the usual (1 – 1/e)-approximation guarantee
  - It might take days…
- What about doing that **on-line**? (e.g., in few milliseconds)
  - Enables on-line analytics for viral marketing, on-line viral ads allocation
  - Need pre-computation and indexing
  - Challenge: enormous number of possible queries! Any possible probability distribution with a state space Z lying on the probability simplex $\triangle^{Z-1}$
- Interesting problem: already several follow-ups (see VLDB'15)

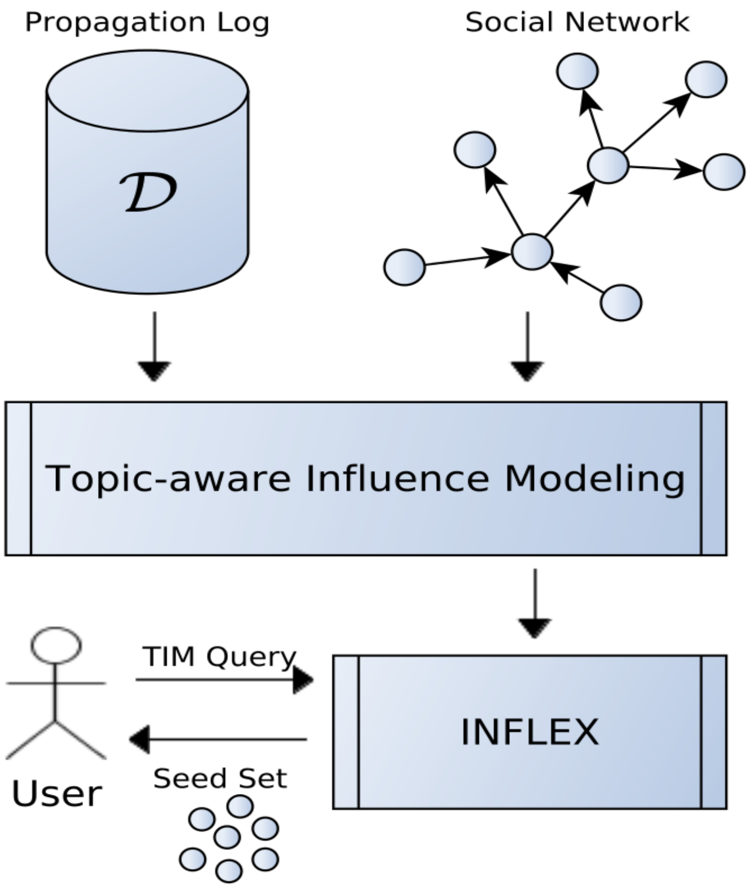
Our idea

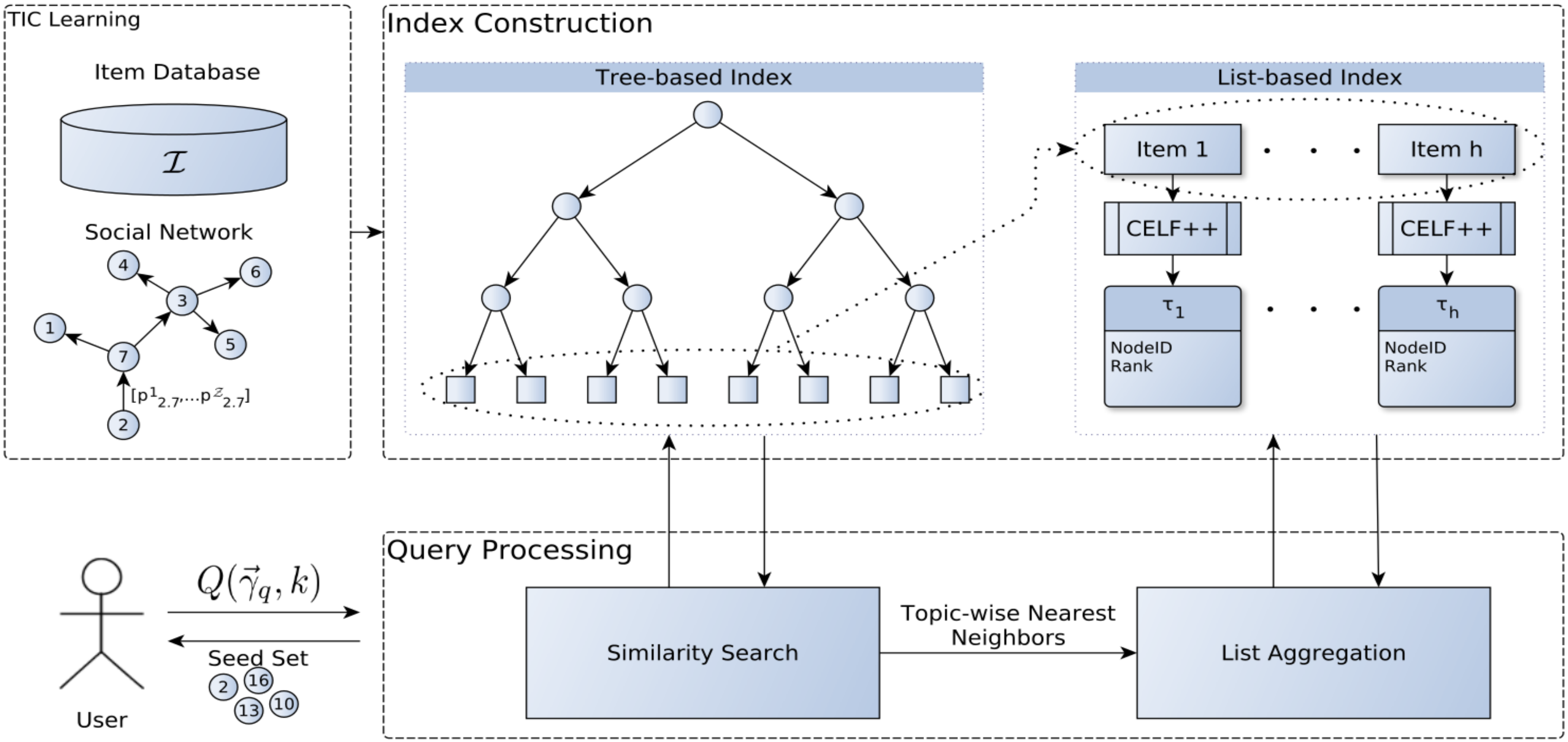
- Similar items are likely to interest similar users:
  - Similar peer influence probabilities
  - Similar influence propagation patterns



INFLEX

Index over pre-computed solutions of a limited number of TIM queries.





Index Construction

- Sample a set of items from the topic space or from the database (off-line step 1)
- extract influential users for the selected index items (off-line step 2)
- index the topic-distributions and the seed node lists (off-line step 3)

Query Processing

- For a given query item, find topic-wise nearest neighbors of the query item in the index (on-line step 1)
- aggregate their pre-computed lists of influential users w.r.t. topic-wise similarity (on-line step 2)

Off-line step 1: index items selection

- Space-based selection: equi-distantly positioned topic distributions on the probability simplex

  (+) Fair coverage of the simplex

  (-) Disregards the available workload

- Data-driven selection: catalog of items learnt from the log of past propagations

  (+) Future queries likely to follow past distributions

  (-) Sparsity issues due to skewed topic distributions in the data

need the best of both approaches... ⟶ Simplex Sampling

Off-line step 1: Simplex Sampling

1. Learn a generative model
   – Estimate the Dirichlet distribution that maximizes the log-likelihood of the available workload

2. Generate many points
   – a large sample with a good simplex coverage

3. Cluster the points
   – applying (Bregman) K-means++ on the sample

4. Use centroids as the items to build the index

5. Number of index items?
   – trade-off between accuracy and efficiency

Off-line step 2: build influential users lists

- extract influential users for the selected index items
- just use any efficient Influence Maximization algorithm

Off-line step 3: build the index

- Bregman Ball Trees[1,2]
  - Hierarchical space partition based on convex Bregman Balls

$$B_f(\mu, R) = \{i \in \mathcal{H} \mid d_f(i, \mu) \leq R\}$$

  - Where the distance is Kullback-Leibler Divergence $D_{KL}(\vec{\gamma}_i \| \vec{\gamma}_q) = \sum_{z=1}^{Z} \gamma_i^z \log \frac{\gamma_i^z}{\gamma_q^z}$
  - Branching: done by means of Bregman k-Means++
  - Branching factor: Mixture of Gaussians clustering to find the number of children of a node

[1] Cayton, *"Fast Nearest Neighbor Retrieval for Bregman Divergences"* ICML 2008
[2] Nielsen et al., *"Tailored Bregman Ball Trees for Effective Nearest Neighbors"* EuroCG 2009

On-line step 1: similarity search

- neither k-NN nor range-search
- dynamic similarity search with Anderson-Darling test
  - If we have close neighbors, we need few of them
  - If there are no very close neighbors, we need more of them to have a more reliable approximation
- If an almost exact match found: directly return its seed set

On-line step 2: rank aggregation

- Combine the rankings of topic-wise nearest neighbors into one "consensus" ranking based on their similarity to the query item
- Compared several rank aggregation methods (see paper)
- In the end we use Weighted Copeland Aggregation
  - It minimizes the number of pairwise disagreements
  - Weight based on KL-divergence from the query item

Adding topic-awarness: summary

- Topic-aware propagation models are important because
  - Influence, expertise, trust are all topic-dependent concepts
  - Real-world applications require topic-aware models
- Extension of IC and LT to their topic-aware versions (TIC, TLT)
  - Nice, elegant, simple, with good properties
  - too many parameters
- AIR model uses topics as proxy and achieves drastic reduction of the number of parameters
- Topic-aware propagation models open the door to new interesting problems: e.g.,
  - On-line TIM Queries (EDBT'14)
  - *How to design viral items?* (see Barbieri and Bonchi, SDM'14)
- Another way of cutting down the number of parameters is to study influence at the community level…

Part II:
Cascades & Communities

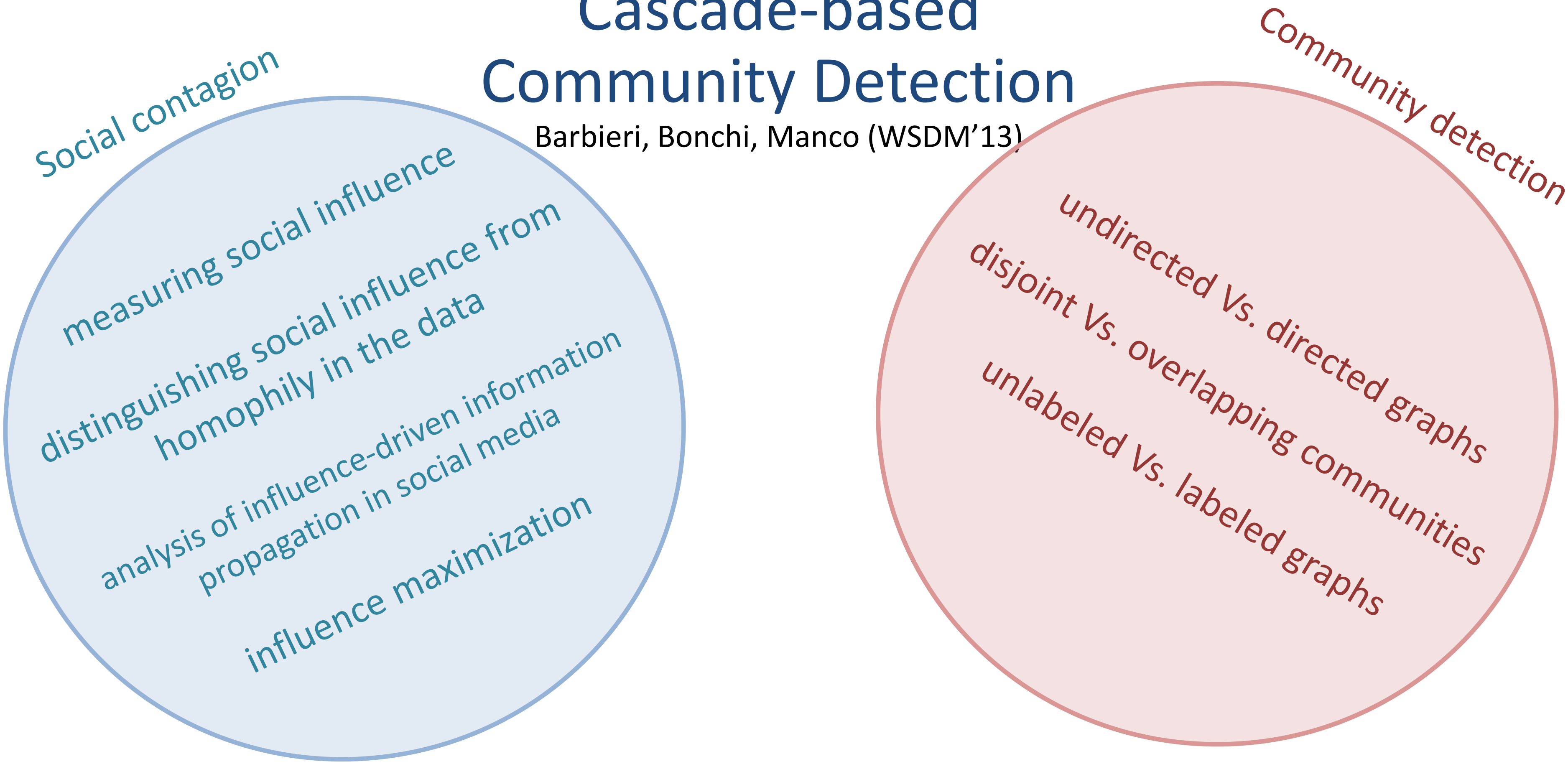
N. Barbieri, F. Bonchi, G. Manco
"Cascade-based Community Detection" (WSDM 2013)

N. Barbieri, F. Bonchi, G. Manco
"Influence-based Network-oblivious Community Detection" (ICDM 2013)

Y. Mehmood, N. Barbieri, F. Bonchi, A. Ukkonen
"CSI: Community-level Social Influence analysis" (ECML-PKDD 2013)

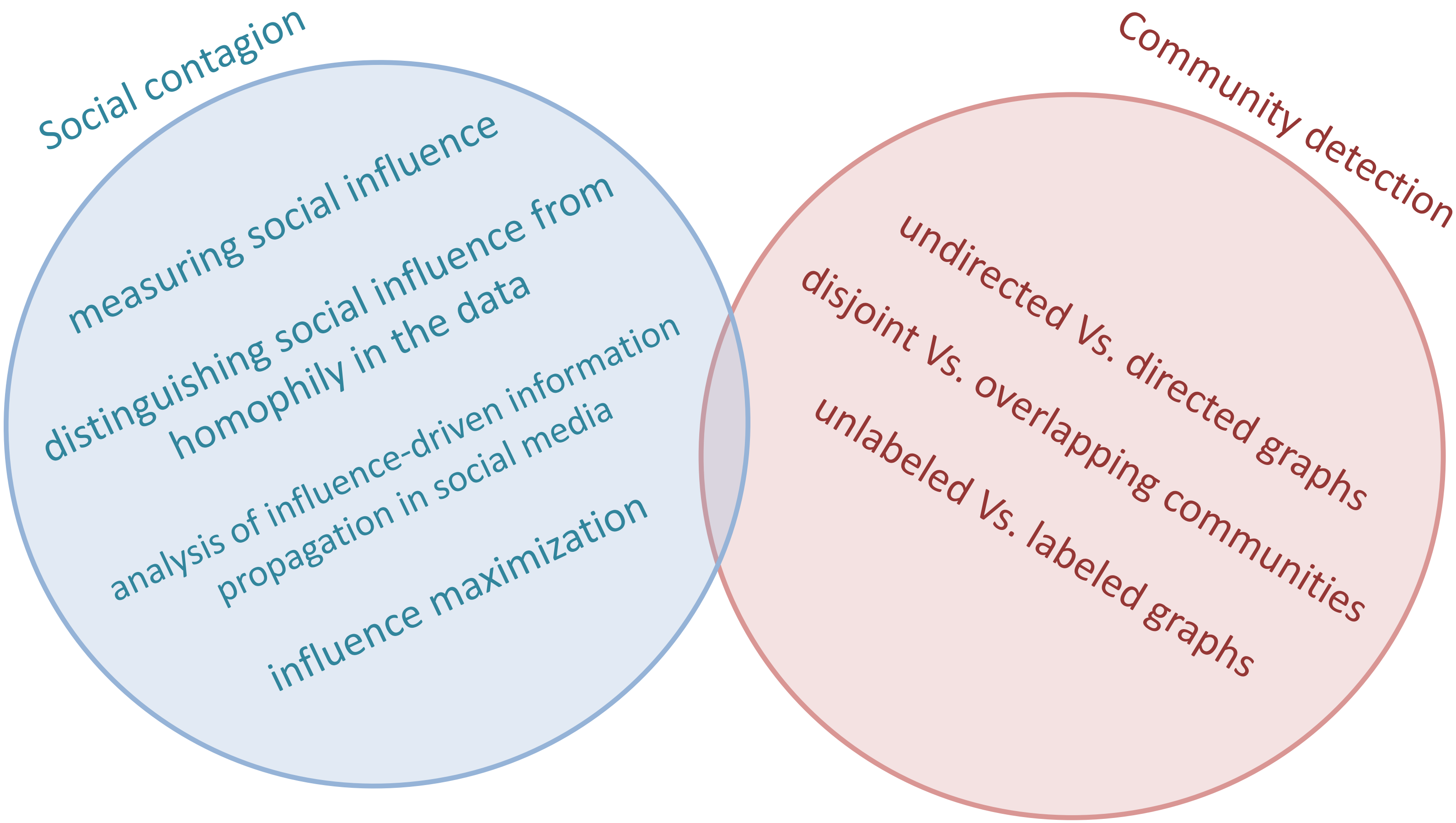
Cascade-based Community Detection

Barbieri, Bonchi, Manco (WSDM'13)



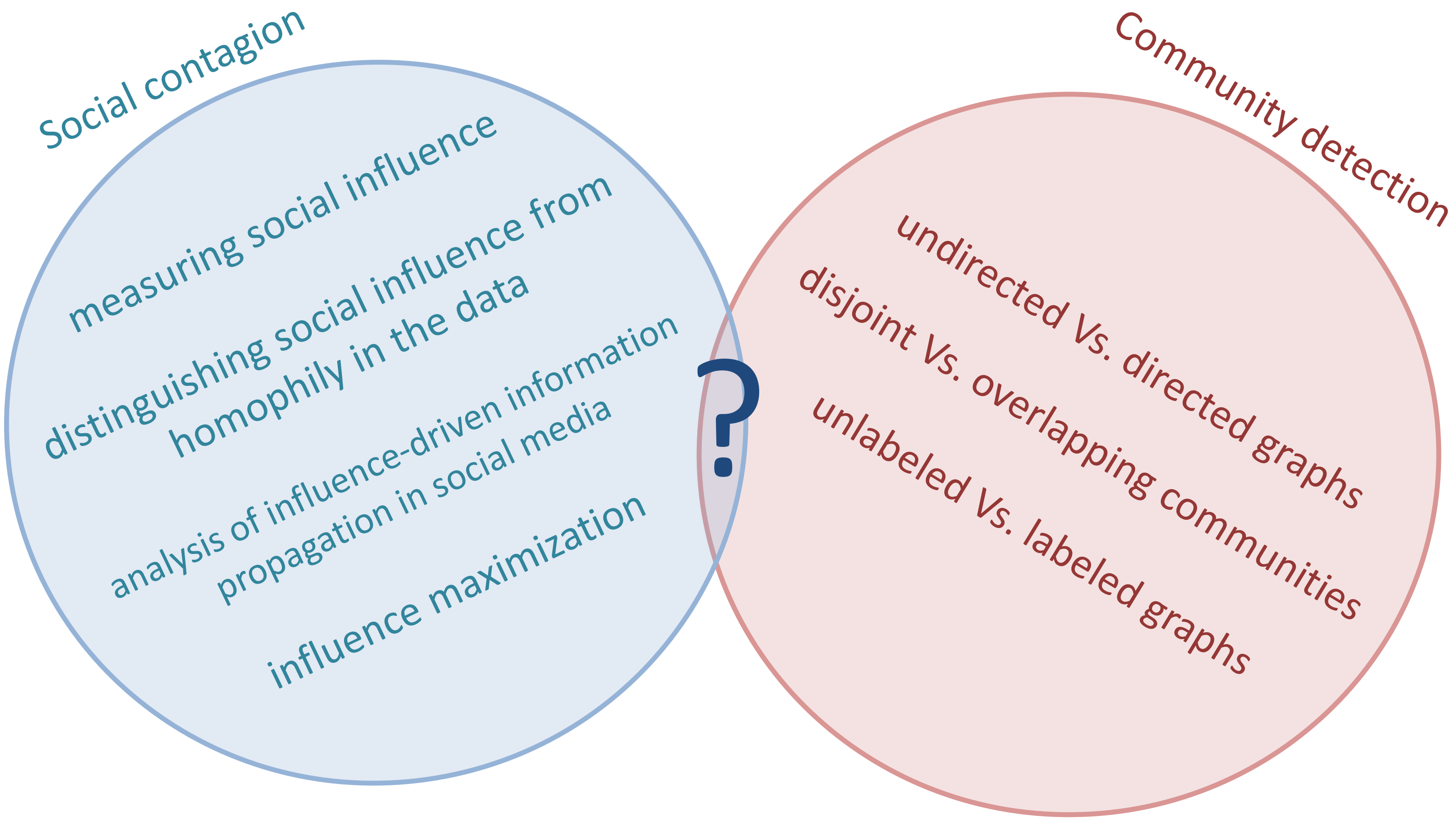
Cascade-based Community Detection

Barbieri, Bonchi, Manco (WSDM'13)



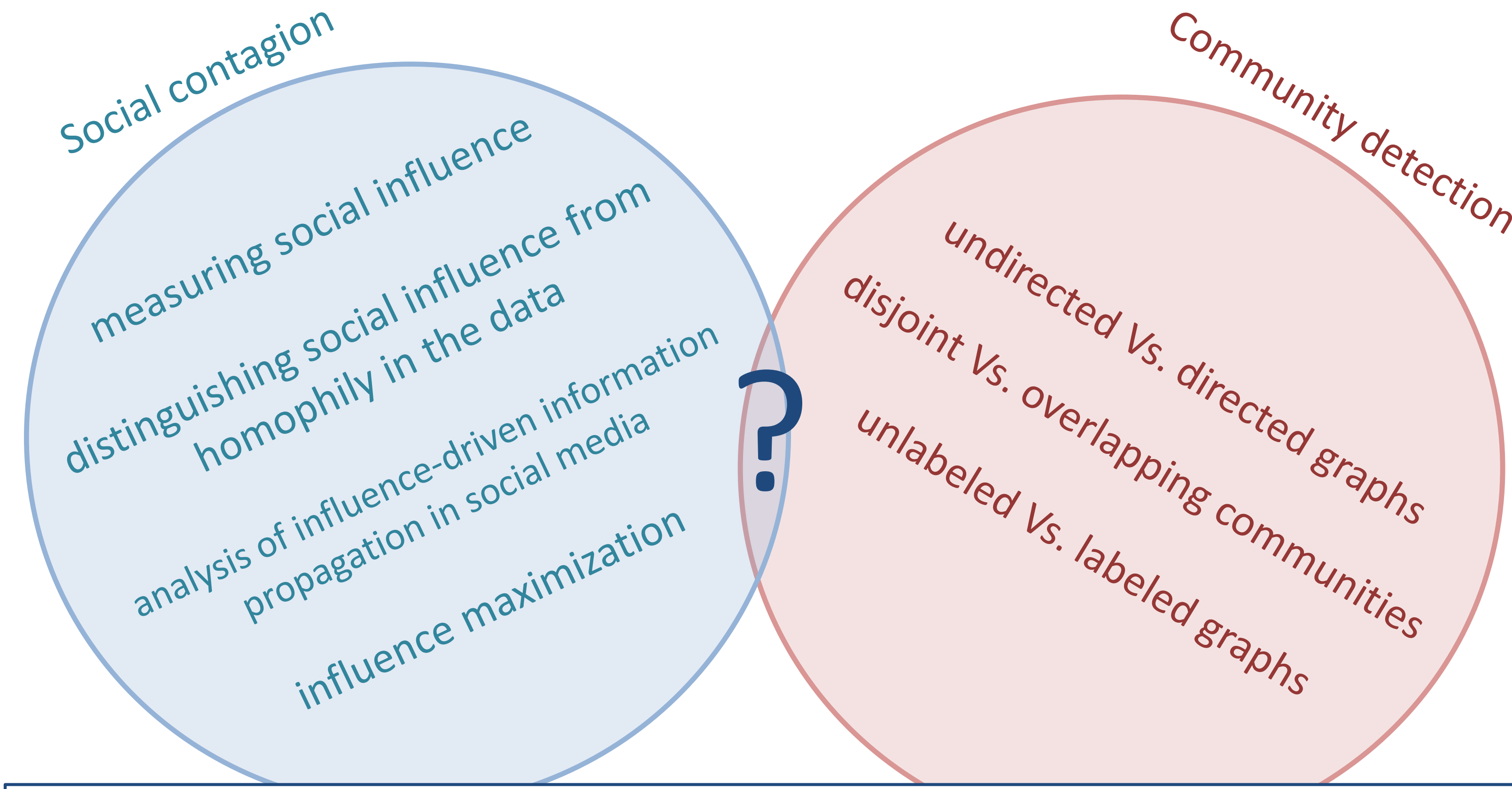
Cascade-based Community Detection

Barbieri, Bonchi, Manco (WSDM'13)



Cascade-based Community Detection

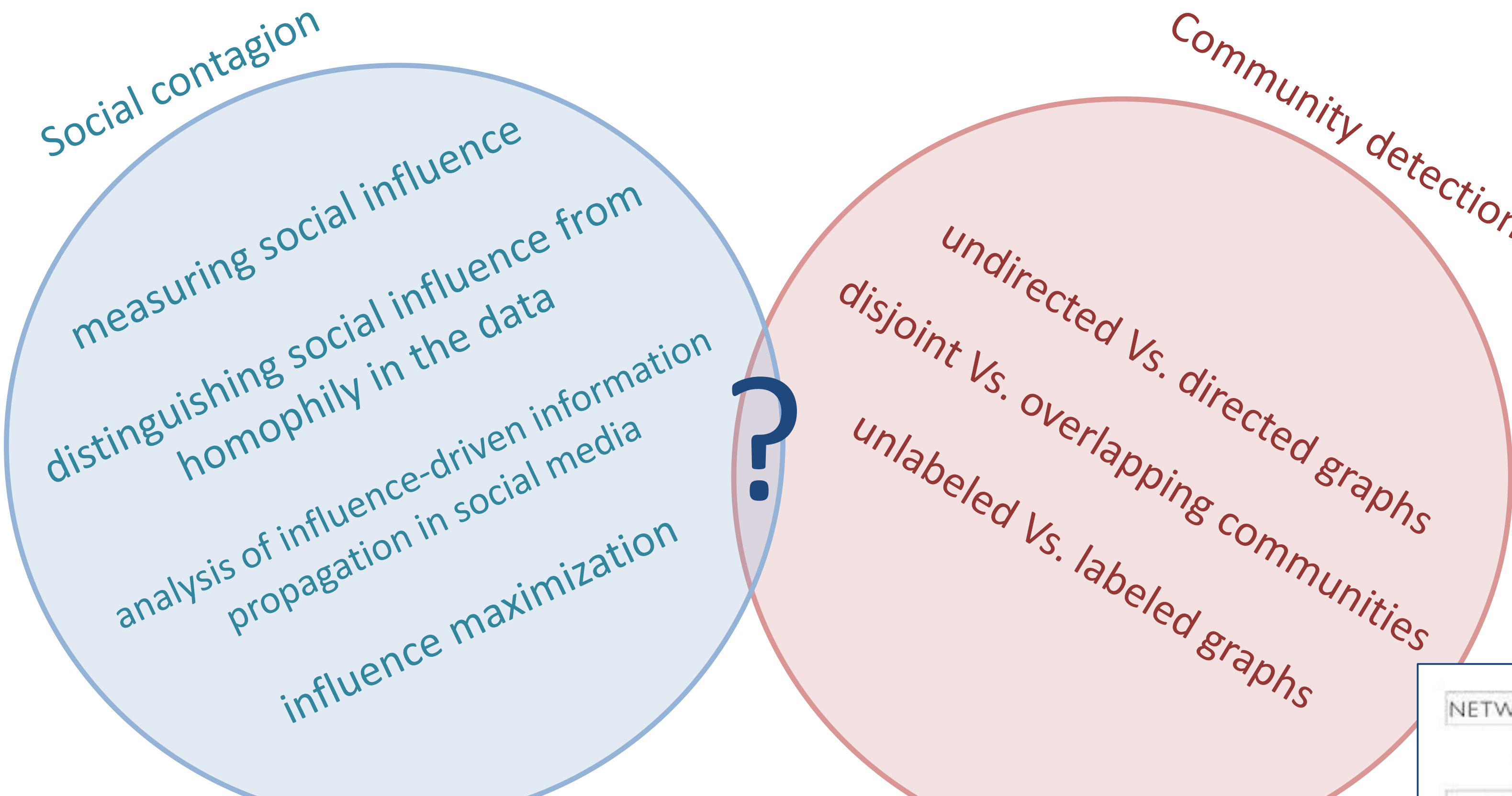
Barbieri, Bonchi, Manco (WSDM'13)



Individuals tend to adopt the behavior of their social peers, so that cascades happen first locally, within close-knit communities, and become global "viral" phenomena only when they are able cross the boundaries of these densely connected clusters of people.

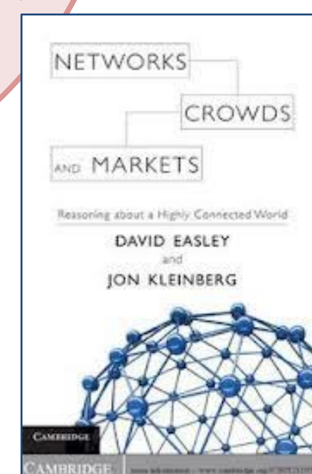
Cascade-based Community Detection

Barbieri, Bonchi, Manco (WSDM'13)



"…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

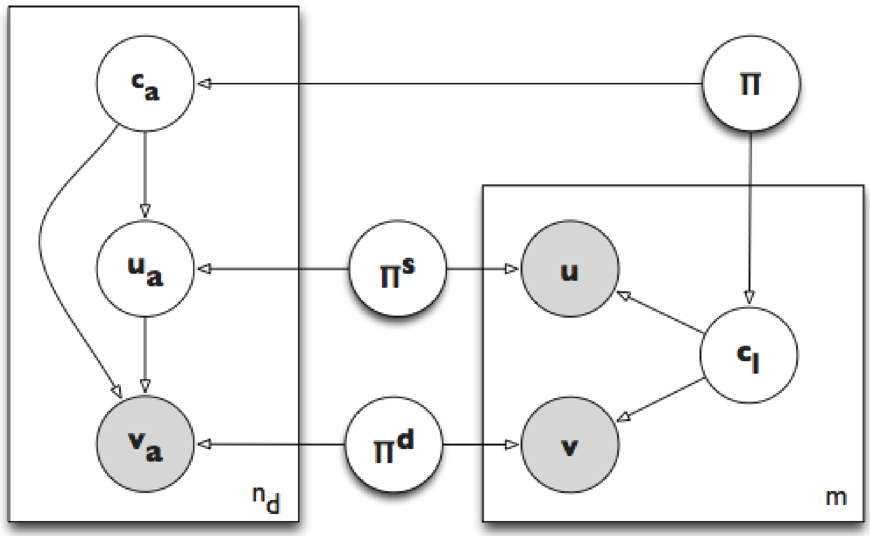
If I'm actively involved in a community I'm followed, and I tweet

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)



each observed action is explained by 3 priors:

the probability Π to observe an action in a community

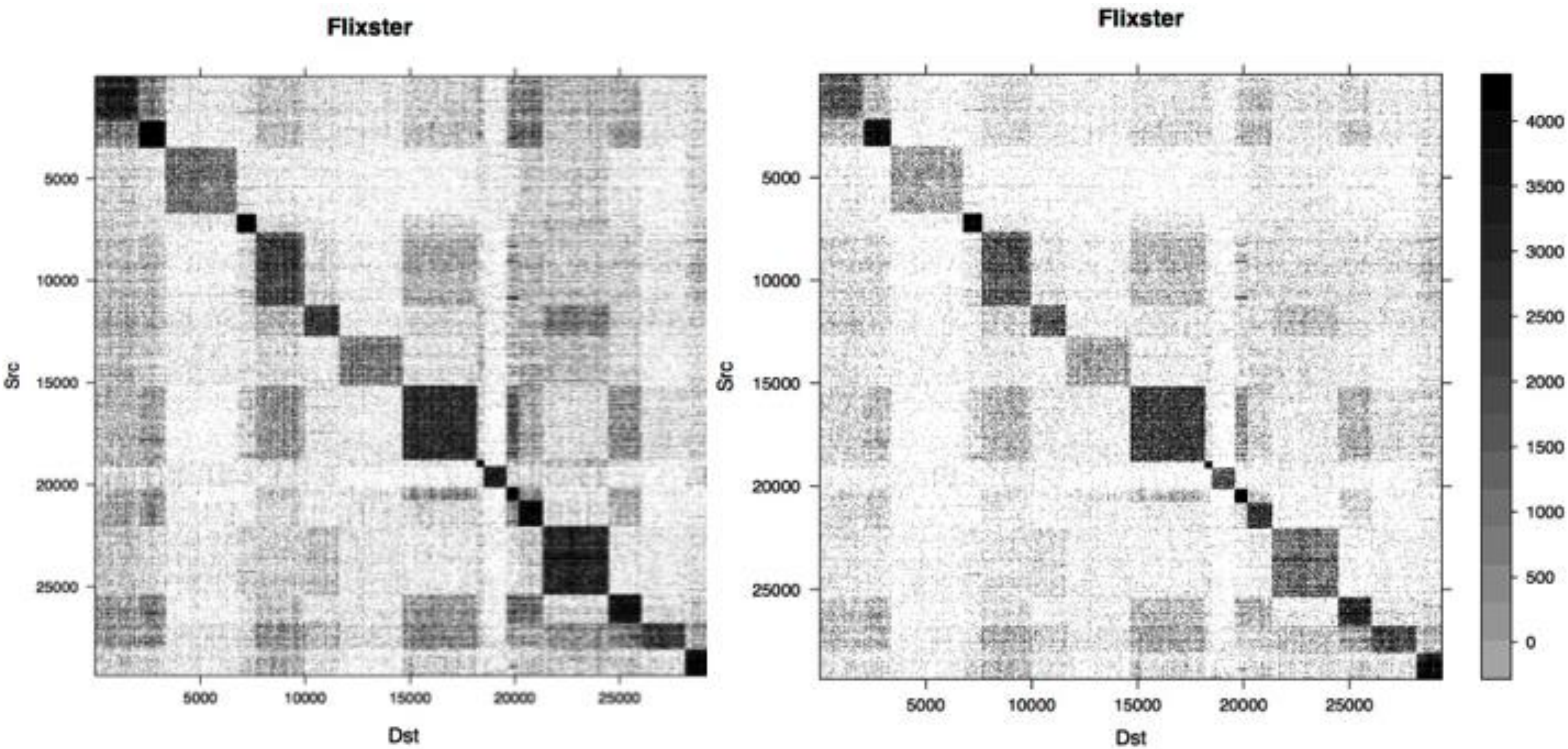
the level of active Π^s of each user in each community

the level of passive Π^d interest of each user in each community

(learning the model parameters: see paper for details)

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

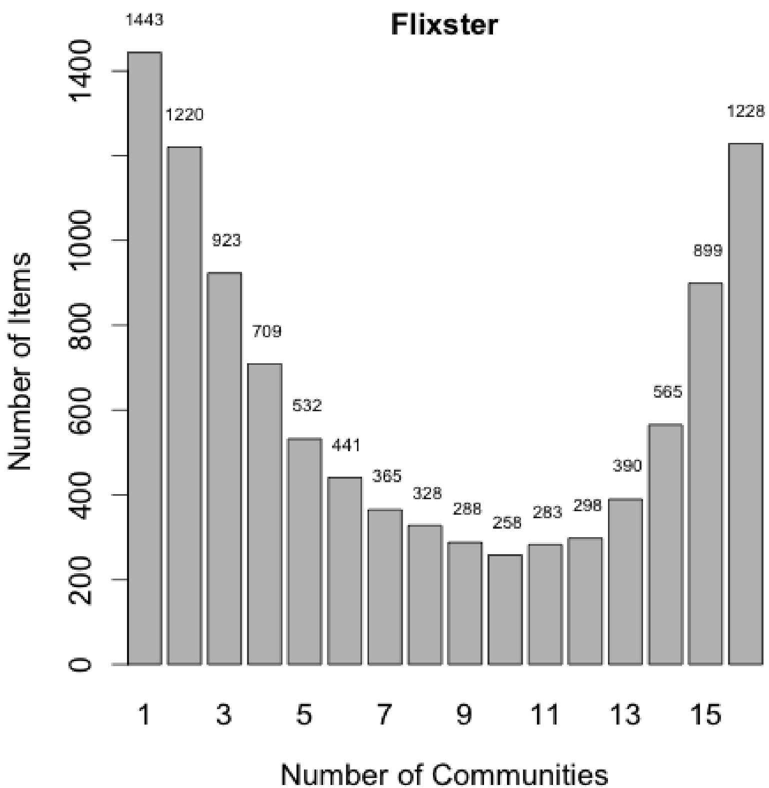


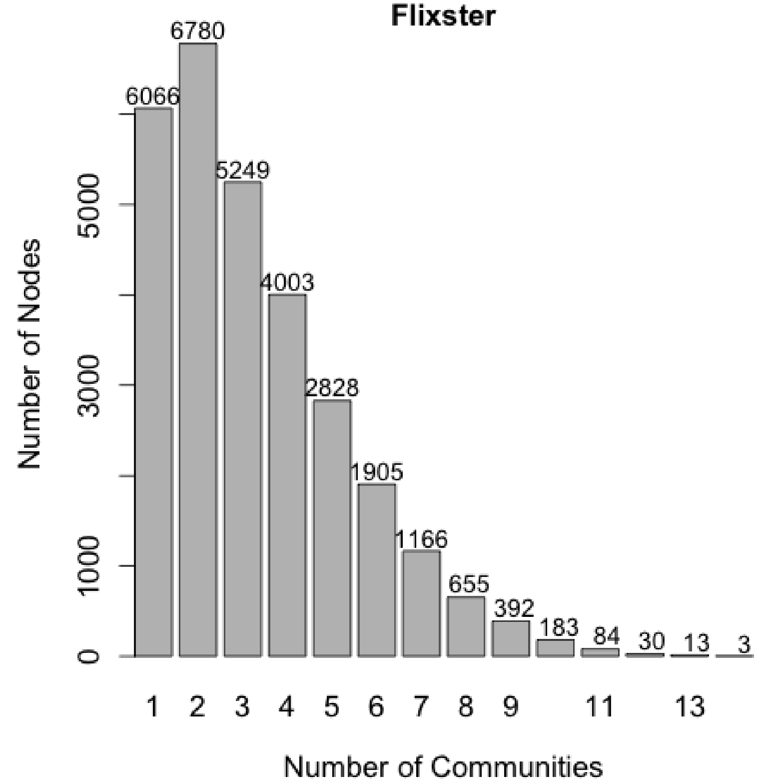
Characterizing the communities

In how many communities users and items tend to participate?

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

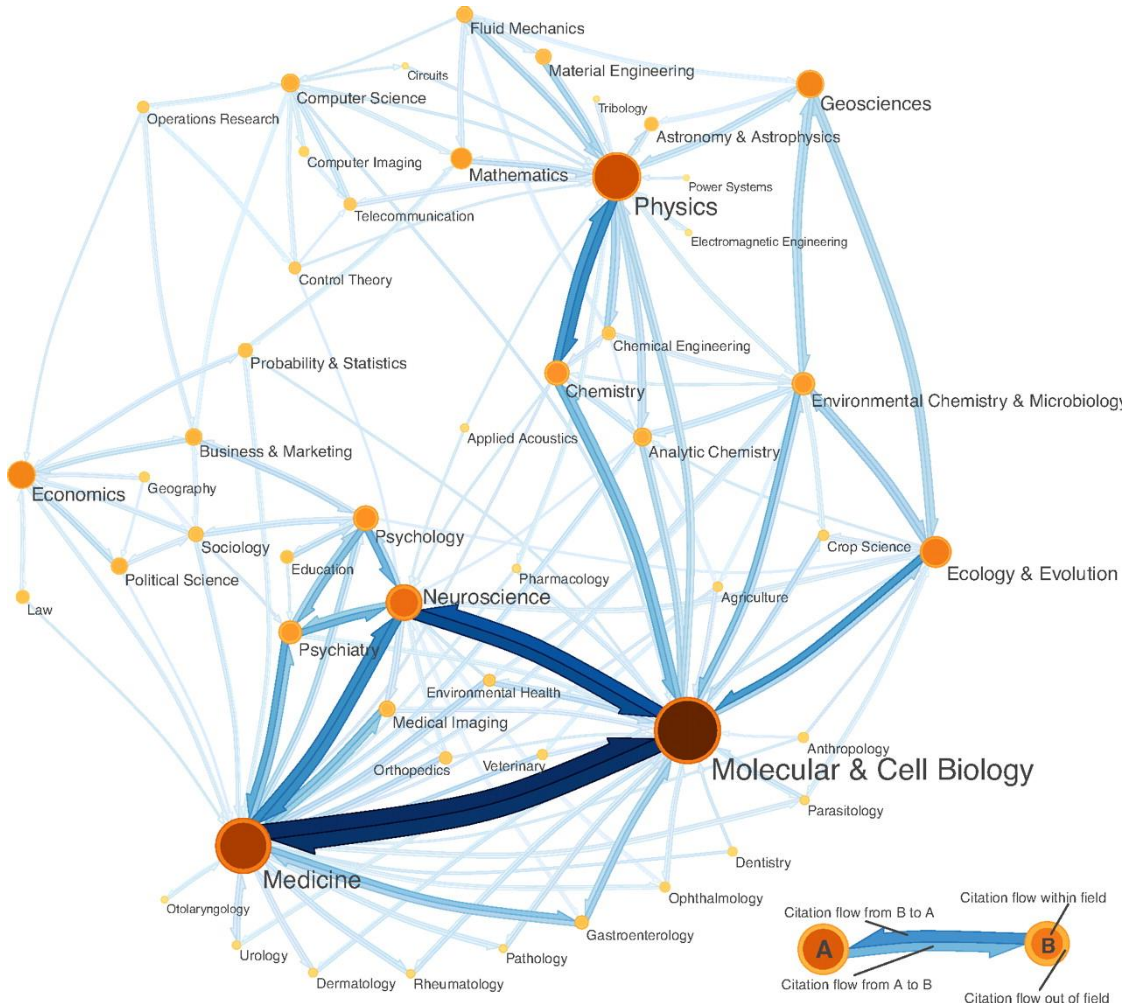
$$\eta_{u,a,k}(\Theta) = P(z_a^k, w_a^u | a \in \mathbb{D}, \Theta)$$





CSI: Community-level Social Influence analysis

Y. Mehmood, N. Barbieri, F. Bonchi, A. Ukkonen (ECML-PKDD'13)

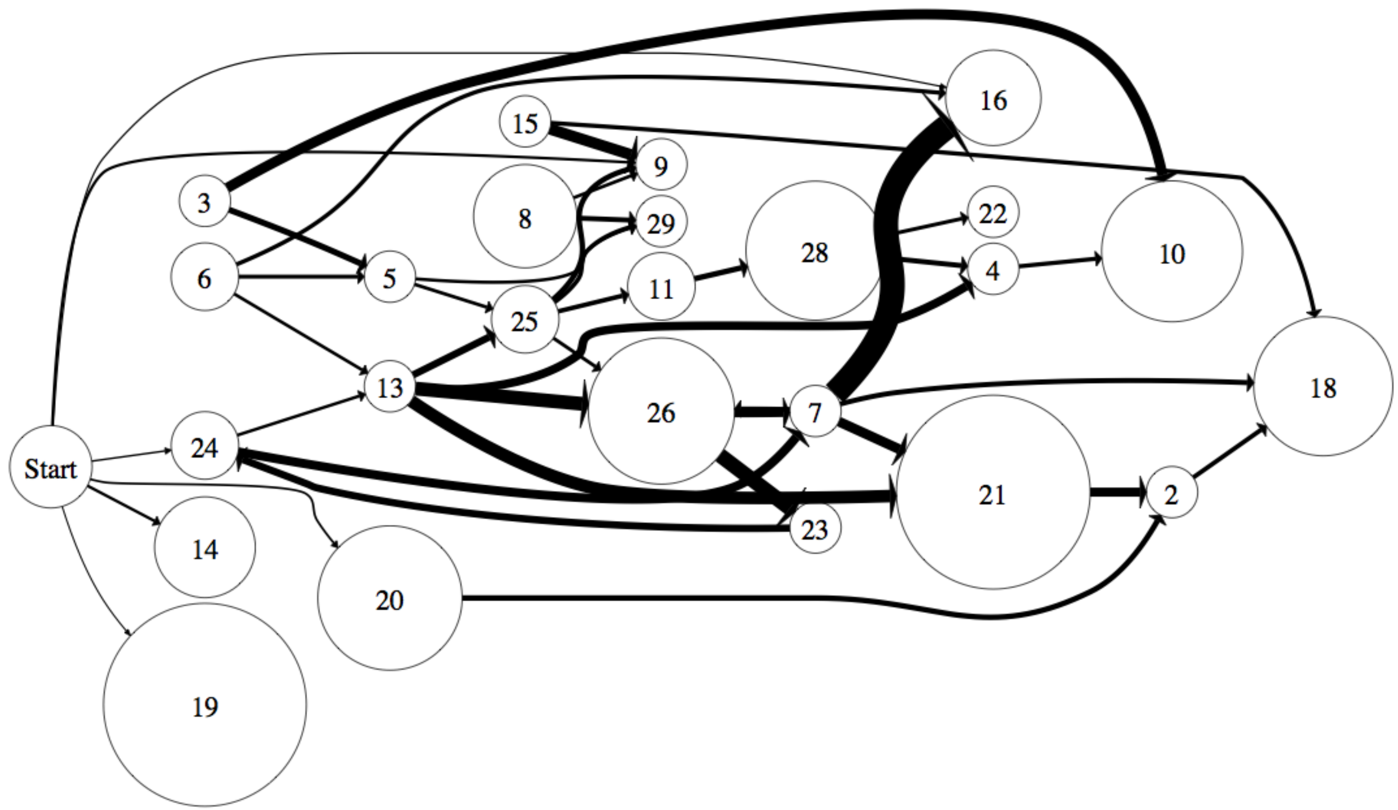


Rosvall M , and Bergstrom C T PNAS 2008;105:1118-1123

A network perspective on modularity, ARCS 2012

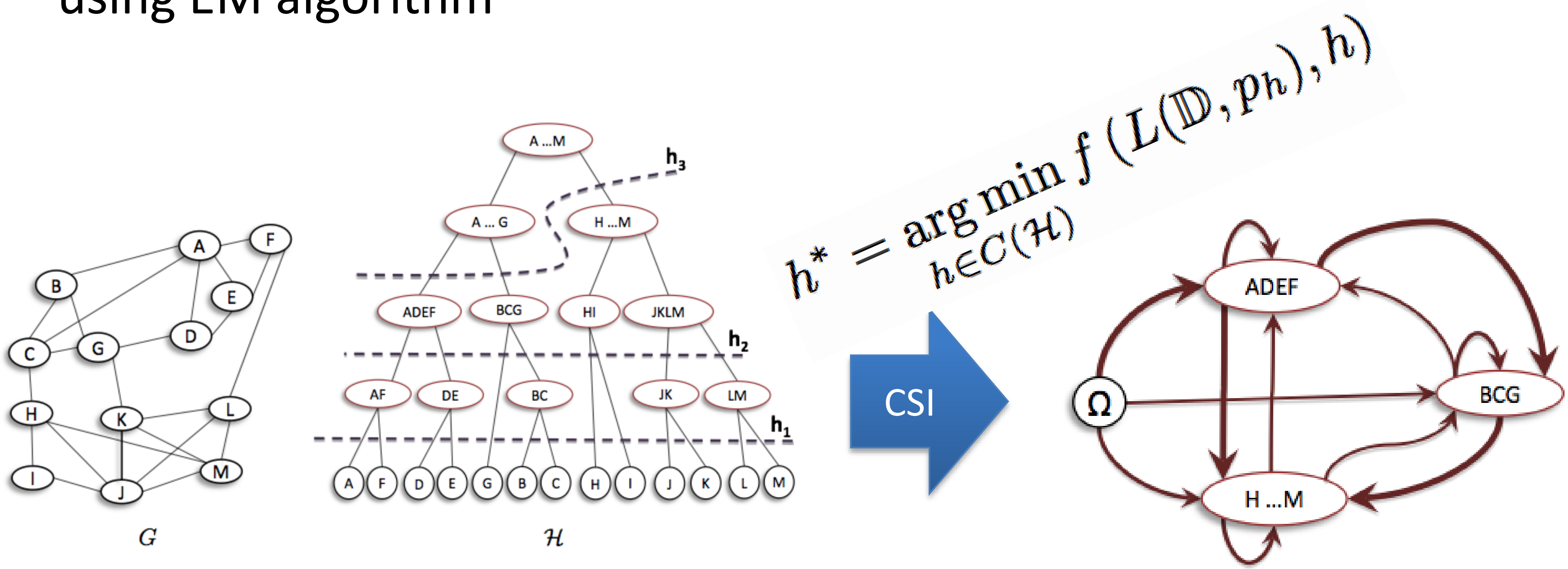
CSI: Community-level Social Influence analysis

[Problem] From a hierarchical partitioning of a social network, find a set of communities + the strength of influence between the communities that better explain the given log of past propagations



CSI Model

- Given past propagations and a hierarchical partitioning of the network CSI finds a model that is a balance between the likelihood (in terms of explaining data) and complexity (in terms of summarization)
- We extend *Saito et al.* approach of learning influence probabilities using EM algorithm



$$h^* = \arg\min_{h \in C(\mathcal{H})} f\left(L(\mathbb{D}, p_h), h\right)$$

$\mathbb{D}$

Ω	A	B	M	Ω	B	C	G	K	……	Ω	A	L	M	C
0	40	45	50	0	10	15	23	47	……	0	20	48	53	97

Influence-based Network-oblivious Community Detection

"Community detection without the network"

Barbieri, Bonchi, Manco (ICDM 2013)

- Why studying community detection without the network?
  - The social graph migh not be available
  - Communication might occur over multiple networks (e-mail, telephone, Facebook, Twitter, Skype or WhatsApp). By being network oblivious we also solve the problem of community detection over multiple networks.

- Possible approaches:
  - [Clustering] Apply normal clustering to the users, where the cascades in which they participate are used as features.
  - [Network reconstruction] First try to reconstruct the unobserved social network from the cascades (methods exist in the literature), then apply some community detection algorithm to the reconstructed network.

Influence-based Network-oblivious Community Detection

"Community detection without the network"

Barbieri, Bonchi, Manco (ICDM 2013)

- Our approach:
  - assumes that item adoptions are governed by un underlying stochastic diffusion process over the unobserved social network, such diffusion model is based on community-level influence.
  - by fitting the model parameters to the user activity log we learn the community membership and the influence level of each user in each community.

- We study two community-level influence diffusion models:
  - Community-level IC model (discrete time)
  - Community-rate model (modeling delays in activation time)

Cascades & Communities: summary

- There is a clear interplay between the modular structure of social networks and the propagation of information
- This can be exploited in different directions
  - Use cascades for finding better communities
  - Exploit the community structure to model cascades and for a coarser social influence analysis → less parameters, overfitting avoidance
  - Not only identify communities, but also infer the type of communities (e.g., social or topical) and the specific roles played by different users in a community [see our KDD'14 paper*]
  - Build applications [see our KDD'14 paper*]

- Exciting and almost unexplored topic at the overlap of two well studied areas
  - Plenty of room for impactful research

* Barbieri, Bonchi, Manco *"Who to Follow and Why: Link Prediction with Explanations"* [KDD'14]

Concluding remarks

Mining information propagation data: many interesting problems

- Modeling cascades
  - Which propagation model is more accurate in modeling the real world?
  - Predicting size of cascades early on
  - Topic-aware models
  - Interplay between information propagation and network evolution (e.g., KDD'13)*

- Social influence
  - Distinguishing social influence from homophily and other factors of correlation
  - Measuring social influence at the link level
  - Measuring social influence at the community level
  - Topic-aware versions of the above problems
  - Streaming and distributed computation (Big Data)
  - Privacy-preserving methods

*Weng et al. *"The Role of Information Diffusion in the Evolution of Social Networks"* (KDD'13)

Mining information propagation data: many interesting problems

- Influence maximization
  - Scalability
  - Quality
  - Direct mining approaches (e.g., VLDB'12)*
  - On-line topic-aware influence maximization queries
  - Competitive viral marketing
  - Revenue maximization and other variants
  - Close the gap with the real world

- Cascades and communities

*Goyal, Bonchi, Lakshmanan *"A Data-Based Approach to Social Influence Maximization"* (VLDB'12)

Social advertising

- How to do advertising on the web exploiting social influence
- New exciting area at the overlap of viral marketing and classic computational advertising
- Need to define the theoretical foundations*
  - auction model, bidding process, pricing model
  - ad allocation mechanism and its properties (e.g., fairness, envy-freeness, etc.)
  - measures of performance
- Many technical challenges*
  - on-line nature, scalability
  - many players with different objectives
  - competitiveness
  - limited attention of the users

*Lu, Bonchi, Goyal, Lakshmanan *"The Bang for the Buck: Fair Competitive Viral Marketing from the Host Perspective"* (KDD'13)

*Aslay, Lu, Bonchi, Goyal, Lakshmanan *"Viral Marketing Meets Social Advertising: Ad Allocation with Minimum Regret "* (VLDB'15)

- Big Data revolution is happening

- The digital traces we all leave on social media are a big part of such revolution.

- Mining such wealth of data might enable:
  - Better understanding of human behaviour
  - Data-driven confirmation/rejection of existing socal theories
  - Better intervention during crises and emergencies
  - Citizens engagement and participation in the social and political life
  - New web applications
  - Viral marketing and new types of social advertising
  - Etc…

- However, classic mining methods are not enough:
  - Semantic richness of the data
  - Expressiveness of the patterns sought
  - Size of the data, scalability
  - Dinamicity and streaming nature of the data
  - Etc…

- New algorithms and methods are needed!

Thank you!
Questions?

@FrancescoBonchi
www.francescobonchi.com
francesco.bonchi@isi.it

Thank you!
Questions?

@FrancescoBonchi
www.francescobonchi.com
francesco.bonchi@isi.it