On Information Propagation, Social Influence, and Communities

@FrancescoBonchi
www.francescobonchi.com
francesco.bonchi@isi.it
Plan of the talk

- Intro: social influence and influence maximization
- Learning the strength of social influence
- Adding topics
- Cascades & communities
- Summary of research problems + an exciting new direction
The Spread of Obesity in a Large Social Network over 32 Years


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 4500 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
• 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**

<table>
<thead>
<tr>
<th>Input:</th>
<th>$G, k, \sigma_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output:</td>
<td>seed set $S$</td>
</tr>
</tbody>
</table>

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 $(1 - \frac{1}{e}) > 63\%$ of the number of nodes that any size-$k$ set could activate.

**Theorem**: The $(1 - \frac{1}{e})$ 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)*
Influence Maximization algorithms

Much work has been done following Kempe et al. mostly devoted 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

Social graph

Propagation log

Learn probabilities

Seed set

Algorithm 1: Greedy
Input: $G, k, \sigma$
Output: seed set $S$
1: $S \leftarrow \emptyset$
2: while $|S| < k$ do
3: select $u = \arg \max_{v \in V \setminus S} \sigma(v)$
4: $S \leftarrow S \cup \{u\}$
Part I:
Learning the strength of social influence
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

<table>
<thead>
<tr>
<th>Action</th>
<th>User</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>u_{12}</td>
<td>1</td>
</tr>
<tr>
<td>a</td>
<td>u_{45}</td>
<td>2</td>
</tr>
<tr>
<td>a</td>
<td>u_{32}</td>
<td>3</td>
</tr>
<tr>
<td>a</td>
<td>u_{76}</td>
<td>8</td>
</tr>
<tr>
<td>b</td>
<td>u_{32}</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>u_{45}</td>
<td>3</td>
</tr>
<tr>
<td>b</td>
<td>u_{98}</td>
<td>7</td>
</tr>
</tbody>
</table>

Action a:
Learning influence strength: EM method
K. Saito, R. Nakano, and M. Kimura
“Prediction of information diffusion probabilities for independent cascade model” (KES’08)

Given the social graph and the log of past propagations, learn influence probability for each link using EM in order to maximize the likelihood of the IC model:

$$\log L(\mathcal{D} \mid p) = \sum_{\alpha \in \mathcal{D}} \log L_\alpha(p)$$

The likelihood of individual trace is:

$$L_\alpha(p) = \prod_{v \in \mathcal{V}} \left[ 1 - \prod_{u \in F^+_\alpha,v} (1 - p(u, v)) \right] \cdot \left[ \prod_{u \in \bar{F}^-_\alpha,v} (1 - p(u, v)) \right]$$

+ elegant and principled
- it assumes the propagation in the propagation log are structured as if they were generated by the IC model: real-world data is different!
- not so obvious to scale
Learning influence strength: frequentist approach

A. Goyal, F. Bonchi, L. V. S. Lakshmanan
“Learning Influence Probabilities In Social Networks” (WSDM’10)

**Static Models:** probabilities are static and do not change over time.

**Bernoulli:** $p_{vu} = \frac{A_{v2u}}{A_v}$

**Jaccard:** $p_{vu} = \frac{A_{v2u}}{A_v|u}$

**Continuous Time (CT) Models:** probabilities decay exponentially in time

$p_{uv}^t = p_{uv}^0 \exp \left( -\frac{t - t_v}{\tau_{uv}} \right)$

Not incremental, hence very expensive to apply on large datasets.

**Discrete Time (CT) Models:** Active neighbor $u$ of $v$ remains contagious in $[t, t+\tau(u,v)]$, has constant influence prob $p(u,v)$ in the interval and 0 outside. **Monotone, submodular, and incremental!**
Comparison of Static, CT and DT models

- Time conscious models are better than the static model
- CT and DT models perform equally well
- Static and DT models are far more efficient compared to CT models because of their **incremental** nature
Learning influence strength: some challenges

Privacy

social graph $G$ proprietary and secret (e.g., Twitter, Facebook)
propagation log $L$ proprietary and secret (e.g., Amazon, Ebay)
different parties hold the two pieces of input

(Tassa & Bonchi – EDBT 2014)

Scalability and streaming

social graph $G$ might not fit in memory
propagation log $L$ potentially huge and streaming

(Kutzkov, Bifet, Bonchi, Gionis - KDD 2013)

Overfitting

we have $|E|$ parameters to learn

(see part II & III of this talk)
Part II: Adding topic-awareness

N. Barbieri, F. Bonchi, G. Manco

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

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

Topic-Aware Independent Cascade (TIC)

$$
p^i_{v,u} = \sum_{z=1}^{K} \gamma_i^z p^z_{v,u}
$$
Learning problem

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

$$\gamma_i^z$$ and $$p_{v,u}^z$$

We devise an EM algorithm for the TIC model

\[
\begin{align*}
\text{E-step} & \quad \text{for all } i \in \mathcal{I} \text{ do} \\
& \quad \quad \text{for all } z = \{1, \cdots, K\} \text{ do} \\
& \quad \quad \quad \quad Q_i(z; \hat{\Theta}) \leftarrow \frac{P(D_i|z; \hat{\Theta})\pi_z}{\sum_{\tilde{z}} P(D_i|\tilde{z}; \hat{\Theta})\pi_{\tilde{z}}}; \\
& \quad \quad \quad \quad \text{for all } (u, v) \in E \text{ do} \\
& \quad \quad \quad \quad \quad R_i^z(u, v; \hat{\Theta}) \leftarrow \frac{p_{v,u}^z}{P_{u,v}^z}; \\
& \quad \quad \text{end} \\
& \quad \text{end} \\
\text{M-step} & \quad \text{for all } z = \{1, \cdots, K\} \text{ do} \\
& \quad \quad \pi_z \leftarrow \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} Q_i(z; \hat{\Theta}); \\
& \quad \quad \text{for all } (u, v) \in E : S_{v,u}^+ \neq \emptyset \text{ do} \\
& \quad \quad \quad p_{v,u}^z \leftarrow \frac{1}{\kappa_{v,u}^+ + \kappa_{v,u}^-} \sum_{i \in S_{v,u}^+} Q_i(z; \hat{\Theta}) R_i^z(u, v; \hat{\Theta}) \\
& \quad \quad \text{end} \\
& \quad \text{end}
\end{align*}
\]

... but:

TIC has a huge number of parameters

#topics( #links + #items)
The AIR propagation model

**Authoritiveness** 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

\[ p^i_{v,u} \]

- \( v \) is authoritative
- \( Z \) is interested
- \( u \) is relevant
- \( i \) is relevant

number of parameters:

\#topics(2 \#nodes + \#items)

[Model definition and learning: see details in the paper (!)]
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 $q$, $\gamma_q$
- 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:

$$Q(\gamma_q, k) = \arg \max_{S \subseteq V, |S| = k} \sigma(S, \gamma_q)$$
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 $\Delta^{Z-1}$

• 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...
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
  \[ D_{KL}(\tilde{\gamma}_i \| \tilde{\gamma}_q) = \sum_{z=1}^{Z} \gamma_i^z \log \frac{\gamma_i^z}{\gamma_q^z} \]
  - 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
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-awareness: 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 III: 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)
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 to cross the boundaries of these densely connected clusters of people.

"...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 $\Pi$ to observe an action in a community.
- The level of active $\Pi^s$ of each user in each community.
- The level of passive $\Pi^d$ 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:

$$\eta_{u,a,k}(\Theta) = P(z^k_a, w^u_a | a \in D, \Theta)$$
[Problem] to 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
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
  - Influence network reconstruction
  - Influence network sparsification
  - 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"
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
  - Incentives models for the users
  - 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 social 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 data/web mining methods are not enough:
  - Semantic richness of the data
  - Expressiveness of the patterns sought
  - Size of the data, scalability
  - Dynamicity and streaming nature of the data
  - Etc...

New algorithms and methods are needed!
Acknowledgments

- Cigdem Aslay
- Ricardo Baeza-Yates
- Nicola Barbieri
- Albert Bifet
- Aris Gionis
- Amit Goyal
- Kostantin Kutzkov
- Laks V.S. Lakshmanan
- Wei Lu
- Giuseppe Manco
- Yasir Mehmood
- Tamir Tassa
- Antti Ukkonen

Thank you!
Questions?

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