Probabilistic Causal Analysis of Social Influence

F. Bonchi\textsuperscript{1}  F. Gullo\textsuperscript{2}  B. Mishra\textsuperscript{3}  D. Ramazzotti\textsuperscript{4}

\textsuperscript{1}ISI Foundation, Italy and Eurecat, Spain, francesco.bonchi@isi.it
\textsuperscript{2}UniCredit, R&D Dept., Italy, gullof@acm.org
\textsuperscript{3}New York University, NY, USA, mishra@nyu.edu
\textsuperscript{4}Stanford University, CA, USA, daniele.ramazzotti@stanford.edu

The 27th ACM International Conference
on Information and Knowledge Management (CIKM 2018)
October 22-26, 2018
Turin, Italy
Social influence: process motivating the actions of a user to induce similar actions from her peers

Mastering the dynamics of social influence is crucial for a variety of applications
- e.g., viral marketing, trust-propagation analysis, personalization, feed ranking, information-propagation analysis

Prior work:
- Estimating the strength of influence in a social network
- Empirically analyzing the effects of social influence
- Distinguishing genuine social influence from homophily and other external factors

Social influence is a genuine causal process: there is no principled causal-theory-based approach to learn social influence from empirical information-propagation data
- We fill this gap!
We devise a principled causal approach to infer social influence from a database of propagation traces. Based on Suppes’ theory of probabilistic causation, our approach outputs a set of causal DAGs describing social influence. Different DAGs imply different communities and different topics.

Major challenges:
- Simpson’s paradox
- Genuine vs. spurious causes

Proposal: a two-step methodology
- **I step**: partitioning the input propagation traces, to get rid of Simpson’s paradox
- **II step**: inferring minimal causal topology (via MLE), to get rid of spurious causes
Outline

- Introduction: motivation, challenges, contributions
- Background
  - information-propagation traces, hierarchical structure, Suppes’ theory
- General (twofold) problem statement
- Problem 1: partitioning the propagation set
  - Problem definition
  - Algorithms
- Problem 2: learning a minimal causal topology
- Experiments
Outline

- Introduction: motivation, challenges, contributions

- Background
  - information-propagation traces, hierarchical structure, Suppes’ theory

- General (twofold) problem statement

- Problem 1: partitioning the propagation set
  - Problem definition
  - Algorithms

- Problem 2: learning a minimal causal topology

- Experiments
Input data

- A (directed) **social graph** \( G = (V, A) \)
- A set \( \mathcal{E} \) of **entities**
- A set \( \mathcal{O} \) of **observations**
  - Triples \( \langle v, \phi, t \rangle \), where \( v \in V, \phi \in \mathcal{E}, t \in \mathbb{N}^+ \)
  - \( \langle v, \phi, t \rangle \in \mathcal{O} \) means: entity \( \phi \) is observed at node \( v \) at time \( t \)
  - Entities cannot be observed multiple times at the same node

Example:

- \( G \): social network (follower-followee relations)
- \( \mathcal{E} \): pieces of multimedia content (posts, photos, videos)
- \( \langle v, \phi, t \rangle \in \mathcal{O} \): multimedia item \( \phi \) enjoyed by user \( v \) at time \( t \)
Observations $\mathcal{O}$ can alternatively be viewed as a database $\mathcal{D}$ of propagation traces, i.e., traces left by entities “flowing” over $G$.

- Propagation trace of an entity $\phi$:
  all observations $\{\langle v, \phi', t \rangle \in \mathcal{O} | \phi' = \phi \}$ involving $\phi$

- $\mathcal{O} \leftrightarrow \mathcal{D} = \{D_\phi | \phi \in \mathcal{E}\}$ of directed acyclic graphs (DAGs)
  - $D_\phi = (V_\phi, A_\phi)$
  - $V_\phi = \{v \in V | \langle v, \phi, t \rangle \in \mathcal{O} \}$
  - $A_\phi = \{(u, v) \in A | \langle u, \phi, t_u \rangle \in \mathcal{O}, \langle v, \phi, t_v \rangle \in \mathcal{O}, t_u < t_v \}$

- No cycles in $D_\phi \in \mathcal{D}$ due to time irreversibility

- All propagations started at time 0 by a dummy node $\Omega \notin V$
Input data: example

<table>
<thead>
<tr>
<th>v</th>
<th>φ</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ω</td>
<td>φ₁</td>
<td>0</td>
</tr>
<tr>
<td>v₂</td>
<td>φ₁</td>
<td>2</td>
</tr>
<tr>
<td>v₃</td>
<td>φ₁</td>
<td>4</td>
</tr>
<tr>
<td>v₄</td>
<td>φ₁</td>
<td>5</td>
</tr>
<tr>
<td>v₅</td>
<td>φ₁</td>
<td>7</td>
</tr>
<tr>
<td>v₆</td>
<td>φ₂</td>
<td>8</td>
</tr>
<tr>
<td>v₇</td>
<td>φ₂</td>
<td>9</td>
</tr>
<tr>
<td>Ω</td>
<td>φ₂</td>
<td>0</td>
</tr>
<tr>
<td>v₂</td>
<td>φ₂</td>
<td>1</td>
</tr>
<tr>
<td>v₁</td>
<td>φ₂</td>
<td>3</td>
</tr>
<tr>
<td>v₅</td>
<td>φ₂</td>
<td>6</td>
</tr>
<tr>
<td>v₇</td>
<td>φ₂</td>
<td>7</td>
</tr>
<tr>
<td>v₆</td>
<td>φ₂</td>
<td>8</td>
</tr>
<tr>
<td>v₃</td>
<td>φ₂</td>
<td>9</td>
</tr>
</tbody>
</table>

F. Bonchi, F. Gullo, B. Mishra, D. Ramazzotti

Probabilistic Causal Analysis of Social Influence
Gupte et al., “Finding hierarchy in directed online social networks”, WWW 2011

Notion of **agony** to reconstruct a proper hierarchical structure of a graph

- Ranking \( r : V \rightarrow \mathbb{N} \)
  - \( r(u) < r(v) \) means \( u \) is “higher” in the hierarchy than \( v \)
  - i.e., the smaller \( r(u) \) is, the more \( u \) is an “early-adopter”
  - \( r(u) < r(v) \) \( \Rightarrow \) \( u \rightarrow v \) is expected \( \Rightarrow \) no “social agony”
  - \( r(u) \geq r(v) \) \( \Rightarrow \) \( u \rightarrow v \) leads to agony: \( u \) has a higher-ranked follower

- Given a graph \( G = (V, A) \) and a ranking \( r \):
  - agony of arc \( (u, v) \): \( \max\{r(u) - r(v) + 1, 0\} \)
  - agony of \( G \): \( a(G, r) = \sum_{(u,v) \in A} \max\{r(u) - r(v) + 1, 0\} \)

- If \( r \) is not provided: look for *a ranking minimizing the agony of \( G \)*

**Agony of a graph** \( G \) is ultimately computed as \( a(G) = \min_r a(G, r) \)
- it takes \( O(|A|^2) \) time [Tatti, ECML PKDD 2014]
DAGs exhibit no agony (just take temporal ordering as a ranking, i.e., \( r(u) = t_u \)).

Merging DAGs may lead to non-zero agony.

E.g., a \( k \)-length cycle (non-overlapping with other cycles) has agony equal to \( k \).

Minimum-agony ranking for \( D_{\phi_1} \cup D_{\phi_2} \):

\[
(v_2:0)(v_1:1)(v_4:2)(v_5:3)(v_7:4)(v_6:5)(v_3:6)
\]

- No agony on all arcs but \( v_3 \rightarrow v_4 \).
- Agony on \( v_3 \rightarrow v_4 \) = length of cycle passing through \( v_3 \) and \( v_4 \) = 5.
Suppes’ probabilistic causation theory

Definition (Prima facie causes [Suppes, 1970])

For any two events $c$ (cause) and $e$ (effect), occurring respectively at times $t_c$ and $t_e$, under the mild assumption that the probabilities $P(c)$ and $P(e)$ of the two events satisfy the condition $0 < P(c), P(e) < 1$, the event $c$ is called a prima facie cause of the event $e$ if it occurs before $e$ and raises the probability of $e$, i.e., $t_c < t_e \land P(e | c) > P(e | \overline{c})$.

Pros:
- Principled causal theory
- Well-established practical effectiveness
- Computationally light (much lighter than other theories, e.g., Judea Pearl’s one)

Cons:
- No notion of spatial proximity
- Prima facie causes may be genuine or spurious: the latter is undesirable
Outline

• Introduction: motivation, challenges, contributions

• Background
  • information-propagation traces, hierarchical structure, Suppes’ theory

• General (twofold) problem statement

• Problem 1: partitioning the propagation set
  • Problem definition
  • Algorithms

• Problem 2: learning a minimal causal topology

• Experiments
General problem statement

Main general goal

Given a database of propagation traces, derive a set of causal DAGs that are well-representative of the social-influence dynamics underlying the input propagations

Desiderata:

1. Get rid of Simpson's paradox
   - if the input data spans multiple causal processes, causal claims may be hidden or misinterpreted

2. Overcome Suppes’ theory cons (especially the spurious-cause one)

We formulate and solve two problems:

- **Agony-Bounded Partitioning**, a combinatorial-optimization problem, for Desideratum 1
- **Minimal Causal Topology**, a learning problem, for Desideratum 2
Introduction: motivation, challenges, contributions

Background
- information-propagation traces, hierarchical structure, Suppes’ theory

General (twofold) problem statement

Problem 1: partitioning the propagation set
  - Problem definition
  - Algorithms

Problem 2: learning a minimal causal topology

Experiments
The **Agony-bounded Partitioning** problem

- **Main requirement:** propagations in a group should be **homogeneous** in terms of their **hierarchical structure**
  
  ⇒ a group of propagations should exhibit **small agony**

- **Further requirements:** groups **limited in size** and with **connected union graphs**

**Problem (Agony-bounded Partitioning)**

Given a set $\mathcal{D}$ of **DAGs** and two positive integers $K, \eta \in \mathbb{N}$, find a partition $\mathcal{D}^* \in \mathcal{P}(\mathcal{D})$ (where $\mathcal{P}(\cdot)$ denotes the set of all partitions of a given set) such that

$$\mathcal{D}^* = \arg\min_{\mathcal{D} \in \mathcal{P}(\mathcal{D})} |\mathcal{D}| \text{ subject to }$$

$$\forall \mathcal{D} \in \mathcal{D} : a(G(\mathcal{D})) \leq \eta, \ |\mathcal{D}| \leq K, \ G(\mathcal{D}) \text{ is weakly-connected}$$

- $G(\mathcal{D})$ is the union graph of all DAGs in $\mathcal{D}$

- $G(\mathcal{D})$ is termed **prima-facie graph**

**Agony-bounded Partitioning** is **NP-hard** (reduction from **Set Cover**)

F. Bonchi, F. Gullo, B. Mishra, D. Ramazzotti
Outline

- Introduction: motivation, challenges, contributions
- Background
  - information-propagation traces, hierarchical structure, Suppes’ theory
- General (twofold) problem statement
- **Problem 1: partitioning the propagation set**
  - Problem definition
  - Algorithms
- Problem 2: learning a minimal causal topology
- Experiments
A simple two-step approximation algorithm

**Algorithm 1 Two-step-Agony-Partitioning**

**Input:** A set $\mathcal{D}$ of DAGs; two positive integers $K$, $\eta$

**Output:** A partition $\mathcal{D}^*$ of $\mathcal{D}$

1: $\mathcal{D}^+ \leftarrow \text{Mine-Valid-DAG-sets}(\mathcal{D}, K, \eta)$

2: $\mathcal{D}^* \leftarrow \text{Greedy-Set-Cover}(\mathcal{D}^+)$

- **Step 1:** frequent-itemset mining
  - DAGs in $\mathcal{D}$ $\equiv$ items
  - support of a DAG set $\equiv a(G(\mathcal{D}))$

- **Step 2:** solving SET COVER on $\mathcal{D}^+ \subseteq 2^\mathcal{D}$ mined in Step 1 gives the optimum

**Theorem**

*Algorithm 1 is a $(\log K)$-approximation for Agony-bounded Partitioning*

- **Pros:** easy-to-implement, quality guarantees
- **Con:** exponential in the size of the input DAG set $\Rightarrow$ really critical!
A more refined sampling-based algorithm

**Algorithm 2 Sampling-Agony-Partitioning**

**Input:** A set $\mathcal{D}$ of DAGs; two positive integers $K$, $\eta$; a real number $\alpha \in (0, 1]$

**Output:** A partition $\mathcal{D}^*$ of $\mathcal{D}$

1. $\mathcal{D}^* \leftarrow \emptyset$, $\mathcal{D}_u \leftarrow \mathcal{D}$
2. while $|\mathcal{D}_u| > 0$ do
3.   $\mathcal{D}_s \leftarrow \emptyset$
4.   while $|\mathcal{D}_s| < \lceil \alpha \times \min\{K, |\mathcal{D}_u|\} \rceil$ do
5.     $\mathcal{D}_s \leftarrow \text{Sample-Maximal-DAG-set}(\mathcal{D}_u, K, \eta)$
6.   $\mathcal{D}^* \leftarrow \mathcal{D}^* \cup \{\mathcal{D}_s\}$, $\mathcal{D}_u \leftarrow \mathcal{D}_u \setminus \mathcal{D}_s$

- $\alpha \in (0, 1]$ trades off between accuracy and efficiency
- Uniform or random maximal frequent-itemset sampling
- Sample-Maximal-DAG-set subroutine: select DAGs from $\mathcal{D}_u$ until
  - $\mathcal{D}_u = \emptyset$, or
  - size $K$ reached, or
  - agony constraint violated

**Theorem**

Algorithm 2 is a $\frac{\log K}{\alpha}$-approximation for **Agony-bounded Partitioning**
Outline

- Introduction: motivation, challenges, contributions
- Background
  - information-propagation traces, hierarchical structure, Suppes’ theory
- General (twofold) problem statement
- Problem 1: partitioning the propagation set
  - Problem definition
  - Algorithms
- Problem 2: learning a minimal causal topology
- Experiments
The **Minimal Causal Topology** problem

- **Main requirement**: remove the *spurious relationships* from every prima-facie graph identified in the previous step
  \[ \Rightarrow \text{select a minimal set of arcs that best explain the input propagations} \]
- **Methodology**:
  1. \( \forall D \in D^* : \text{reconstruct a DAG } G_D(D) \text{ from } G(D) \)
  2. \( \forall G_D(D) : \text{learn its minimal causal topology via (constrained) maximum likelihood estimation (MLE)} \)

**Problem (Minimal Causal Topology)**

*Given a database \( D \) of propagations and a DAG \( G_D(D) = (V_D, A_D) \), find \( A^*_D(D) = \arg \max_{\hat{A}_D \subseteq A_D} f(\hat{A}_D, D) \), where \( f(\hat{A}, D) = LL(D|\hat{A}) - R(\hat{A}) \), \( LL(\cdot) \) is the log-likelihood, and \( R(\cdot) \) is a regularization term.*

- As a likelihood score, we experimented with both BIC and AIC

Even if constrained, **Minimal Causal Topology** is still an MLE **NP**-hard problem
  \[ \Rightarrow \text{we adopt a classic greedy hill-climbing heuristic} \]
Outline

- Introduction: motivation, challenges, contributions
- Background
  - information-propagation traces, hierarchical structure, Suppes’ theory
- (Twofold) Problem Statement
  - Problem 1: partitioning the propagation set
  - Problem 2: learning a minimal causal topology
- Algorithms for Problem 1
- Experiments
Experiments on synthetic data: efficiency

Figure: Synthetic data: **execution time** of the proposed PSC method (milliseconds), by varying the $\alpha$ and $\eta$ parameters and the social graph ($|\mathcal{O}| = 1000$, noise 5%, BIC regularizator)
Experiments on synthetic data: impact of $\alpha$ and $\eta$ on effectiveness

- **accuracy** = $\frac{TP + TN}{TP + TN + FP + FN}$
- **NMI** to measure the similarity between the PSC’s clusters and the ground-truth clusters
- **B**: baseline that performs only Step 1

↓

Table: Synthetic data: **effectiveness** of the proposed PSC method vs. the baseline, **by varying the $\alpha$ and $\eta$ parameters**, on the power-law $\delta = 0.05$ social graph ($|\mathbb{O}| = 1000$, noise 5%, BIC regularizator)

<table>
<thead>
<tr>
<th></th>
<th>$\eta = 0$</th>
<th>$\eta = 1$</th>
<th>$\eta = 3$</th>
<th>$\eta = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>accuracy</strong></td>
<td>PSC</td>
<td>0.979</td>
<td>0.979</td>
<td>0.979</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>0.938</td>
<td>0.938</td>
<td>0.942</td>
</tr>
<tr>
<td><strong>NMI</strong></td>
<td></td>
<td>0.563</td>
<td>0.563</td>
<td>0.563</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\eta = 0$</th>
<th>$\eta = 1$</th>
<th>$\eta = 3$</th>
<th>$\eta = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>accuracy</strong></td>
<td>PSC</td>
<td>0.979</td>
<td>0.978</td>
<td>0.979</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>0.938</td>
<td>0.938</td>
<td>0.941</td>
</tr>
<tr>
<td><strong>NMI</strong></td>
<td></td>
<td>0.563</td>
<td>0.563</td>
<td>0.563</td>
</tr>
</tbody>
</table>
Experiments on synthetic data:
impact of $|\mathcal{O}|$ on effectiveness

Table: Synthetic data: **effectiveness** of the proposed PSC method vs. the baseline, by varying the size $|\mathcal{O}|$ of input observations ($\alpha=0.1$, $\eta=1$, noise 5%, BIC regularizer)

| $|\mathcal{O}|$ | Erdős-Rény | PSC | B | NMI |
|--------------|-------------|-----|---|-----|
| 500          | 0.939       | 0.815 | 0.669 |
| 1000         | 0.932       | 0.767 | 0.662 |
| 5000         | 0.909       | 0.585 | 0.662 |
### Real data

- $|\mathcal{O}|$: number of observations
- $|\mathcal{D}|$: number of propagations/DAGs
- $|V|$ and $|A|$: nodes and arcs of the social graph $G$
- $n_{\text{min}}$, $n_{\text{max}}$, and $n_{\text{avg}}$: min, max, and avg number of nodes in a DAG of $\mathcal{D}$
- $m_{\text{min}}$, $m_{\text{max}}$, and $m_{\text{avg}}$: min, max, and avg number of arcs in a DAG of $\mathcal{D}$

| $|\mathcal{O}|$ | $|\mathcal{D}|$ | $|V|$ | $|A|$ | $n_{\text{min}}$ | $n_{\text{max}}$ | $n_{\text{avg}}$ | $m_{\text{min}}$ | $m_{\text{max}}$ | $m_{\text{avg}}$ |
|---|---|---|---|---|---|---|---|---|---|
| Last.fm | 1208640 | 51495 | 1372 | 14708 | 6 | 472 | 24 | 5 | 2704 | 39 |
| Twitter | 580141 | 8888 | 28185 | 1636451 | 12 | 13547 | 66 | 11 | 240153 | 347 |
| Flixster | 6529012 | 11659 | 29357 | 425228 | 14 | 16129 | 561 | 13 | 85165 | 1561 |
Experiments on real data: spread prediction

- No ground-truth $\Rightarrow$ we resort to a **spread-prediction task**
  - predict the nodes activated through an information-propagation process

- We use the Goyal et al.’s propagation model defined in “**A data-based approach to social influence maximization**”, VLDB 2011
  - learns a spread-prediction model from a graph and a set of propagations

- We randomly split propagations into training set and test set (70%-30%), and learn the Goyal et al.’s model on the former

- Graph: our causal structure vs. the whole input social graph

- We predict spread (by 10K Monte Carlo simulations) on the test set, and measure accuracy by MSE
Experiments on real data: spread prediction

<table>
<thead>
<tr>
<th>Twitter</th>
<th>Mean Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>1e+08</td>
</tr>
<tr>
<td>BIC</td>
<td>1e+08</td>
</tr>
<tr>
<td>Social Network</td>
<td>2e+08</td>
</tr>
</tbody>
</table>

Figure: Spread-prediction performance of the proposed PSC method (equipped with BIC or AIC regularizator) vs. a baseline that considers the whole social graph ($\alpha = 0.2$, $\eta = 5$)
Conclusion

• We tackle the problem of deriving causal DAGs that are well-representative of the social-influence dynamics underlying an input database of propagation traces.

• We devise a principled two-step methodology that is based on Suppes’ probabilistic-causation theory.

• The first step of the methodology aims at partitioning the input set of propagations, mainly to get rid of the Simpson’s paradox, while the second step derives the ultimate minimal causal topology via constrained MLE.

• Experiments on synthetic data attest to the high accuracy of the proposed method in detecting ground-truth causal structures, while experiments on real data show that our method performs well in a task of spread prediction.
Thanks!
Experiments on synthetic data: impact of regularizer on effectiveness

Table: Synthetic data: **effectiveness** of the proposed PSC method **by varying the regularizer**, i.e., **BIC vs. AIC** ($|\mathcal{O}| = 1000$, noise 5%, power-law $\delta = 0.05$ social graph)

<table>
<thead>
<tr>
<th>$\alpha = 0.1$</th>
<th>$\eta = 0$</th>
<th>$\eta = 1$</th>
<th>$\eta = 3$</th>
<th>$\eta = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BIC</td>
<td>AIC</td>
<td>BIC</td>
<td>AIC</td>
</tr>
<tr>
<td><strong>accuracy</strong></td>
<td>0.979</td>
<td>0.971</td>
<td>0.979</td>
<td>0.971</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\alpha = 0.25$</th>
<th>$\eta = 0$</th>
<th>$\eta = 1$</th>
<th>$\eta = 3$</th>
<th>$\eta = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BIC</td>
<td>AIC</td>
<td>BIC</td>
<td>AIC</td>
</tr>
<tr>
<td><strong>accuracy</strong></td>
<td>0.979</td>
<td>0.971</td>
<td>0.978</td>
<td>0.971</td>
</tr>
</tbody>
</table>
Experiments on synthetic data: impact of noise level on effectiveness

Table: Synthetic data: **effectiveness** of the proposed PSC method vs. the baseline, **by varying the noise level** ($\alpha = 0.1$, $\eta = 1$, $|\mathcal{O}| = 1000$, BIC regularizer)

<table>
<thead>
<tr>
<th>Power-law $\delta = 0.1$</th>
<th>no noise</th>
<th>noise 5%</th>
<th>noise 10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSC</td>
<td>0.967</td>
<td>0.965</td>
<td>0.964</td>
</tr>
<tr>
<td>B</td>
<td>0.887</td>
<td>0.882</td>
<td>0.878</td>
</tr>
<tr>
<td>NMI</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
</tr>
</tbody>
</table>
Experiments on real data: spread prediction

Figure: Spread-prediction performance of the proposed PSC method (equipped with BIC or AIC regularizator) vs. a baseline that considers the whole social graph ($\alpha = 0.2$, $\eta = 5$)