Probabilistic Causal Analysis of Social Influence

F. Bonchi[1] F. Gullo[2] B. Mishra[3] D. Ramazzotti[4]

[1]ISI Foundation, Italy and Eurecat, Spain, francesco.bonchi@isi.it

[2]UniCredit, R&D Dept., Italy, gullof@acm.org

[3]New York University, NY, USA, mishra@nyu.edu

[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

Motivation

- **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!**

Challenges and Contributions

- We devise a principled **causal approach** to infer social influence from a database of **propagation traces**
  - Based on **Suppes**' theory of **probabilistic causation**
  - Output: a set of **causal DAGs** describing social influence
  - Different DAGs $\Rightarrow$ different communities, 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 $\mathbb{O}$ of **observations**
  - Triples $\langle v, \phi, t \rangle$, where $v \in V, \ \ \phi \in \mathcal{E}, \ \ t \in \mathbb{N}^+$
  - $\langle v, \phi, t \rangle \in \mathbb{O}$ means: entity ϕ 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)
- $\langle v, \phi, t \rangle \in \mathbb{O}$: multimedia item ϕ enjoyed by user v at time t

Input data: information-propagation traces

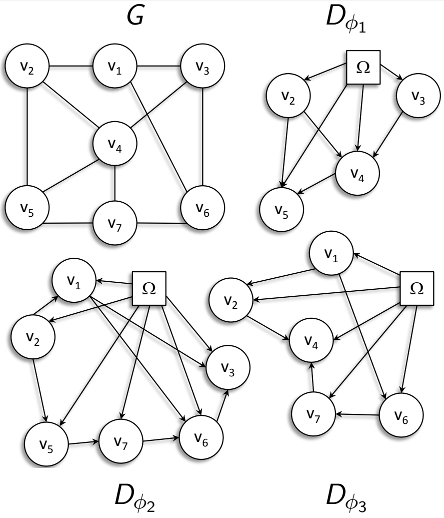
Observations $\mathbb{O}$ can alternatively be viewed as a database $\mathbb{D}$ of **propagation traces**, i.e., traces left by entities "flowing" over G

- Propagation trace of an entity ϕ:
  all observations $\{\langle v, \phi', t\rangle \in \mathbb{O} \mid \phi' = \phi\}$ involving ϕ
- $\mathbb{O} \Leftrightarrow \mathbb{D} = \{D_\phi \mid \phi \in \mathcal{E}\}$ of **directed acyclic graphs** (DAGs)
  - $D_\phi = (V_\phi, A_\phi)$
  - $V_\phi = \{v \in V \mid \langle v, \phi, t\rangle \in \mathbb{O}\}$
  - $A_\phi = \{(u, v) \in A \mid \langle u, \phi, t_u\rangle \in \mathbb{O}, \langle v, \phi, t_v\rangle \in \mathbb{O}, t_u < t_v\}$
- No cycles in $D_\phi \in \mathbb{D}$ due to **time irreversibility**
- All propagations started at time 0 by a **dummy node** $\Omega \notin V$

Input data: example

$\mathbb{D}$

v	ϕ	t
Ω	ϕ_1	0
v_2	ϕ_1	2
v_3	ϕ_1	4
v_4	ϕ_1	5
v_5	ϕ_1	7
Ω	ϕ_2	0
v_2	ϕ_2	1
v_1	ϕ_2	3
v_5	ϕ_2	6
v_7	ϕ_2	7
v_6	ϕ_2	8
v_3	ϕ_2	9
Ω	ϕ_3	0
v_1	ϕ_3	1
v_2	ϕ_3	3
v_6	ϕ_3	5
v_7	ϕ_3	7
v_4	ϕ_3	8



Hierarchical structure

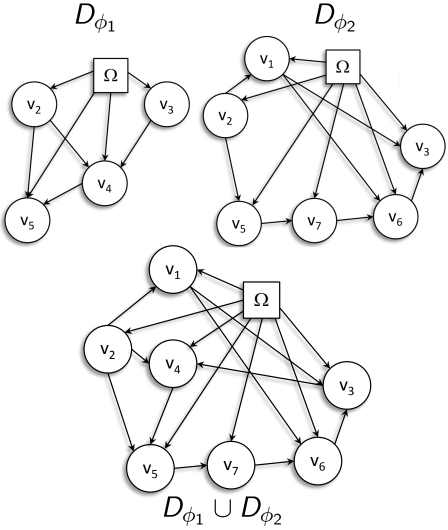
Gupte *et al.*, *"Finding hierarchy in directed online social networks"*, WWW 2011

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

- Ranking $r : V \to \mathbb{N}$
  - $r(u) < r(v)$ means u is "higher" in the hierarchy than v
  - i.e., the smaller $r(u)$ is, the more u is an "early-adopter"
  - $r(u) < r(v) \Rightarrow u \to v$ is expected $\Rightarrow$ no "**social agony**"
  - $r(u) \geq r(v) \Rightarrow u \to v$ 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 $\mathcal{O}(|A|^2)$ time [Tatti, ECML PKDD 2014]

Hierarchical structure: example



- 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 $\mathcal{P}(c)$ and $\mathcal{P}(e)$ of the two events satisfy the condition $0 < \mathcal{P}(c), \mathcal{P}(e) < 1$, 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., $t_c < t_e \ \wedge \ \mathcal{P}(e \mid c) > \mathcal{P}(e \mid \overline{c})$.

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

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 AGONY-BOUNDED PARTITIONING problem

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

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

$$\mathbf{D}^* \quad = \quad \text{argmin}_{\mathbf{D} \in \mathcal{P}(\mathbb{D})} |\mathbf{D}| \qquad \textit{subject to}$$
$$\forall \mathcal{D} \in \mathbf{D}: \; a(G(\mathcal{D})) \leq \eta, \quad |\mathcal{D}| \leq K, \quad G(\mathcal{D}) \textit{ is weakly-connected}$$

- $G(\mathcal{D})$ 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)

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 $\mathbb{D}$ of DAGs; two positive integers K, η
Output: A partition $\mathbf{D}^*$ of $\mathbb{D}$

1: $\mathbf{D}^+ \leftarrow$ Mine-Valid-DAG-sets$(\mathbb{D}, K, \eta)$
2: $\mathbf{D}^* \leftarrow$ Greedy-Set-Cover$(\mathbf{D}^+)$

- Step 1: **frequent-itemset mining**
  – DAGs in $\mathbb{D} \equiv$ items
  – support of a DAG set $\equiv a(G(\mathcal{D}))$
- Step 2: solving SET COVER on $\mathbf{D}^+ \subseteq 2^{\mathbb{D}}$ 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 $\mathbb{D}$ of DAGs; two positive integers K, η; a real number $\alpha \in (0, 1]$

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

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

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

  $\Rightarrow$ select a **minimal set of arcs** that **best explain** the input propagations

- Methodology:
  1. $\forall \mathcal{D} \in \mathbf{D}^*$: reconstruct a DAG $G_D(\mathcal{D})$ from $G(\mathcal{D})$
  2. $\forall G_D(\mathcal{D})$: learn its minimal causal topology via (constrained) **maximum likelihood estimation** (**MLE**)

Problem (Minimal Causal Topology)

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

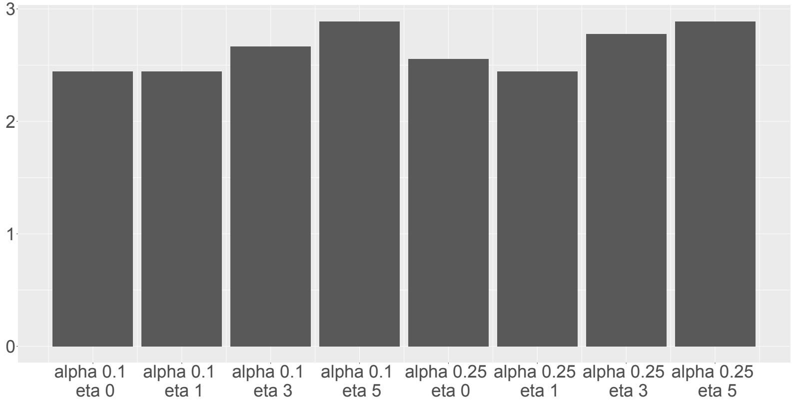
- As a likelihood score, we experimented with both BIC and AIC

Even if constrained, Minimal Causal Topology is still an MLE **NP**-hard problem
$\Rightarrow$ 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



(a) Erdös-Rény social graph

Figure: Synthetic data: **execution time** of the proposed PSC method (milliseconds), by varying the α and η parameters and the social graph ($|\mathbb{O}| = 1000$, noise 5%, BIC regularizator)

Experiments on synthetic data: impact of α and η 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

$\Downarrow$

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

		$\alpha = 0.1$				$\alpha = 0.25$			
		$\eta = 0$	$\eta = 1$	$\eta = 3$	$\eta = 5$	$\eta = 0$	$\eta = 1$	$\eta = 3$	$\eta = 5$
accuracy	PSC	0.979	0.979	0.979	0.98	0.979	0.978	0.979	0.98
	B	0.938	0.938	0.942	0.944	0.938	0.938	0.941	0.945
NMI		0.563	0.563	0.563	0.563	0.563	0.563	0.563	0.563

Experiments on synthetic data: impact of $|\mathbb{O}|$ on effectiveness

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

		Erdös-Rény		
		$\|\mathbb{O}\| = 500$	$\|\mathbb{O}\| = 1000$	$\|\mathbb{O}\| = 5000$
accuracy	PSC	0.939	0.932	0.909
	B	0.815	0.767	0.585
NMI		0.669	0.662	0.662

Real data

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

	$\lvert\mathbb{O}\rvert$	$\lvert\mathbb{D}\rvert$	$\lvert V\rvert$	$\lvert A\rvert$	n_{min}	n_{max}	n_{avg}	m_{min}	m_{max}	m_{avg}
Last.fm	1 208 640	51 495	1 372	14 708	6	472	24	5	2 704	39
Twitter	580 141	8 888	28 185	1 636 451	12	13 547	66	11	240 153	347
Flixster	6 529 012	11 659	29 357	425 228	14	16 129	561	13	85 165	1 561

Experiments on real data: spread prediction

- No ground-truth ⇒ 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

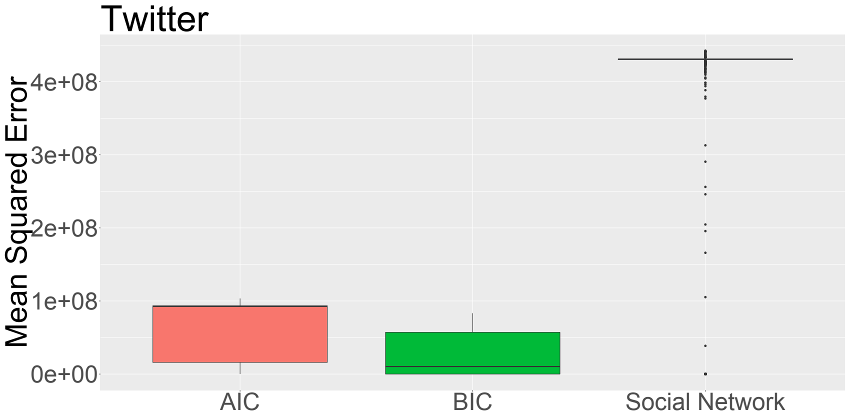


Figure: Spread-prediction performance of the proposed PSC method (equipped with BIC or AIC regularizator) vs. a baseline that considers the whole social graph ($\alpha = 0.2$, $\eta = 5$)

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 regularizator on effectiveness

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

$\alpha = 0.1$

	$\eta = 0$		$\eta = 1$		$\eta = 3$		$\eta = 5$	
	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC
accuracy	0.979	0.971	0.979	0.971	0.979	0.972	0.98	0.973

$\alpha = 0.25$

	$\eta = 0$		$\eta = 1$		$\eta = 3$		$\eta = 5$	
	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC
accuracy	0.979	0.971	0.978	0.971	0.979	0.972	0.98	0.973

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** ($\alpha = 0.1$, $\eta = 1$, $|\mathbb{O}| = 1000$, BIC regularizator)

		Power-law $\delta = 0.1$		
		no noise	noise 5%	noise 10%
accuracy	PSC	0.967	0.965	0.964
	B	0.887	0.882	0.878
NMI		0.63	0.63	0.63

Experiments on real data: spread prediction

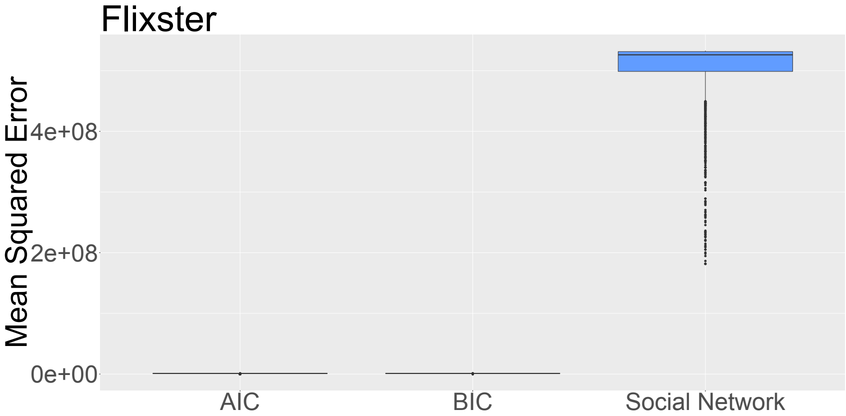


Figure: Spread-prediction performance of the proposed PSC method (equipped with BIC or AIC regularizator) vs. a baseline that considers the whole social graph ($\alpha = 0.2$, $\eta = 5$)