Online Influence Maximization

Siyu Lei, Silviu Maniu, Luyi Mo, Reynold Cheng, Pierre Senellart

KDD 2015, Sydney, Australia
August 13, 2015

Influence Maximization

Important problem in social networks, with applications in marketing, computational advertising

- **objective**: given a promotion budget, maximize the influence spread in the social network (word-of-mouth effect)

  select *k* seeds (influencers) in the social graph, given an influence graph and a propagation model

Influence Maximization

Data model: influence graph *G(V,E,p)*, where

- *V* and *E* and the vertices (users) and edges (follow relations, friendship, etc.) in the social network,
- *p* is a function mapping edges to influence probabilities.

Influence Maximization

Independent cascade model — a discrete time model of propagation:

- at time *0* — activate the seed *s*,
- node *i* activated at time *t* — influence is propagated at *t+1* to neighbours *j* independently with probability *p(i,j)*,
- once a node is activated, it cannot be deactivated or activated again.

Influence Maximization

The independent cascade model is a stochastic process

Influence maximization in this model tries to optimize the expected influence spread, $\sigma(S)$, from a set of seeds S.

Influence Maximization

Influence maximization is computationally hard — two sources of hardness:

- computing *σ(S)* is hard = evaluating probability formulas
- even if we know *σ(S),* computing the influence maximisation is NP-hard (submodular maximization subject to a constraint)

Solutions:

- for computing *σ(S):* Monte Carlo simulations of influence spread
- for solving the influence maximization: greedy approximation algorithm

Multiple algorithms and estimators: CELF, TIM / TIM+

Online Influence Maximization (OIM)

What if we only know the social graph, but still want to maximize influence, with a budget?

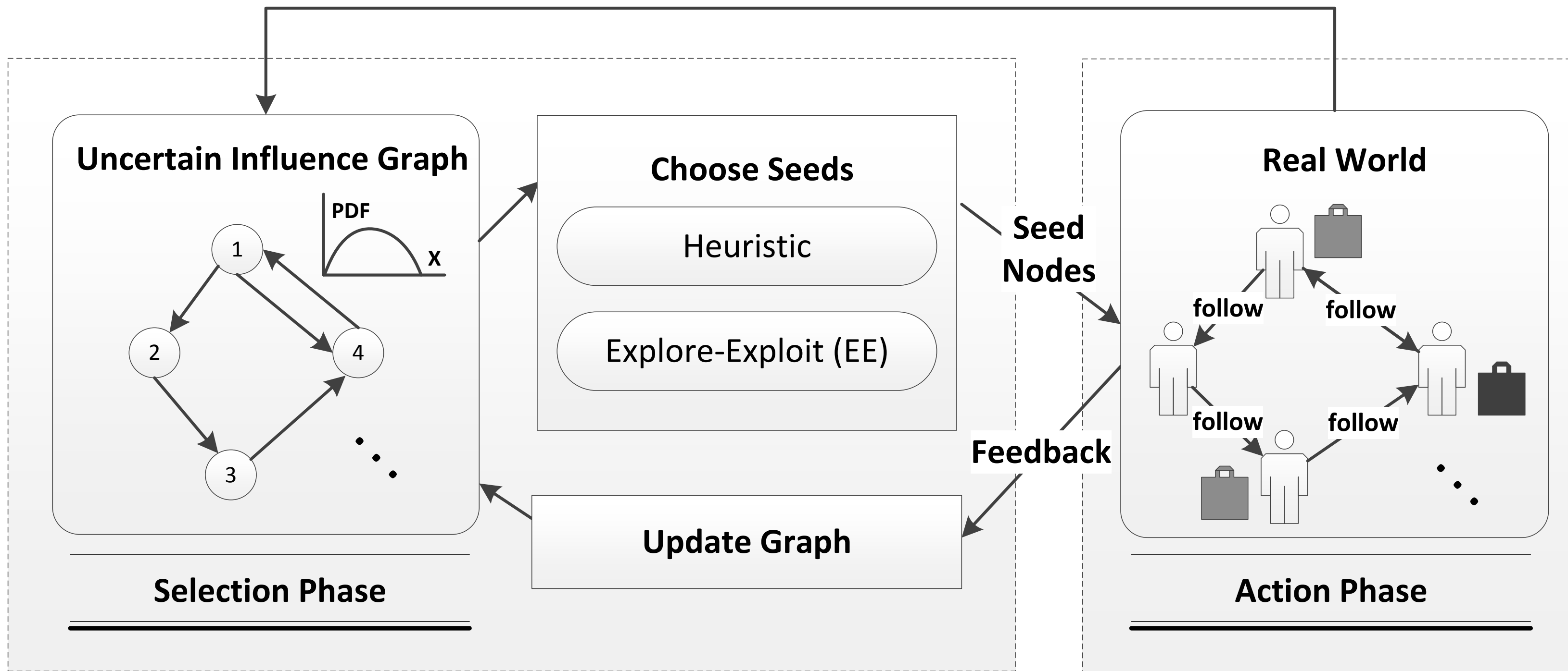
- we need to keep an (uncertain) model of the influence graph
- classic trade-off between exploration (refine the model) and exploitation (use the model to maximize influence)
- lends itself to an iterative process over several rounds (online)

Online Influence Maximization Problem

Maximize the influence spread given a budget of *N* rounds of choosing *k* seeds in the network

- **Contribution**: an online framework — maximization and model refinement over multiple rounds

OIM Framework



Input: # trials N, budget k, uncertain influence graph G
Output: seed nodes $S_n (n = 1 \ldots N)$, activation results A
$A \leftarrow \emptyset$
for $n = 1$ **to** N **do**
  $S_n \leftarrow \texttt{Choose}(G, k)$
  $(A_n, F_n) \leftarrow \texttt{RealWorld}(S_n)$
  $A \leftarrow A \cup A_n$
  $\texttt{Update}(G, F_n)$
return $\{S_n | n = 1 \ldots N\}, A$

OIM Framework

Three ingredients:

- the model of the influence graph
- the explore-exploit strategy (`Choose`)
- after real-world feedback, update of the model (`Update`)

Uncertain Influence Graph

Probabilistic graph model:

- instead of a probability *p(i,j)* on each edge *(i,j)*, we associate it with a distribution of probabilities

$$P(i,j) \sim \mathrm{Beta}(\alpha_{ij}, \beta_{ij})$$

- by default, each edge is associated with a prior probability distribution $\mathrm{Beta}(\alpha, \beta)$

Choose Strategies

The uncertain graph model allows us to explore different assumptions about the graph:

- exploit assumes that the influence probabilities are the expected value of *P(i,j)*

- explore uses either other assumptions about the graph, or uses heuristic strategies (random, max degree, degree discount)

For each branch, the IM algorithm is a black box (CELF, TIM, …) only the input influence graph is different

Choose: Confidence Bound

A classic approach to use other assumptions about the influence graph is the Confidence Bound (CB) algorithm:

- each edge distribution is "moved" by θ standard deviations, and the IM algorithm is executed
- allows to "explore" other "possible influence graphs"
- exploit corresponds to the case where θ is 0

1: **Input:** uncertain influence graph $G = (V, E, P)$, budget k
2: **Output:** seed nodes S with $|S| = k$
3: **for** $e \in E$ **do**
4: $\quad \mu_{ij} \leftarrow \frac{\alpha_{ij}}{\alpha_{ij}+\beta_{ij}}$
5: $\quad \sigma_{ij} \leftarrow \frac{1}{(\alpha_{ij}+\beta_{ij})} \cdot \sqrt{\frac{\alpha_{ij}\beta_{ij}}{(\alpha_{ij}+\beta_{ij}+1)}}$
6: $\quad p_{ij} \leftarrow \mu_{ij} + \theta\sigma_{ij}$
7: $G' \leftarrow G$, with edge probabilities $p_{ij}, \forall (i,j) \in E$
8: $S \leftarrow \mathtt{IM}(G', k)$
9: **return** S

A probabilistic parameter ε allows the choice between different θ values (including 0 for exploit) — similar to ε-greedy

`Choose`: Confidence Bound

Advantages of CB:

- allows the update of ε probabilities for a fixed choice of θ values — Exponentiated Gradient (EG)
- using CB with EG allows a theoretical regret bound for a given choice of (constant) θ values

Input: $\vec{\varphi}$, probability distribution; δ, accuracy parameter; G_n, the gain obtained; j, the index of latest used θ_j; $\mathbf{w}$, a vector of weights; N, the number of trials.
Output: θ
$\gamma \leftarrow \sqrt{\frac{\ln(q/\delta)}{qN}}$, $\tau \leftarrow \frac{4q\gamma}{3+\gamma}$, $\lambda \leftarrow \frac{\tau}{2q}$
for $i = 1$ **to** q **do**
    $w_i \leftarrow w_i \times \exp\left(\lambda \times \frac{G_n \times \mathbb{I}[i=j]+\gamma}{\varphi_i}\right)$
for $i = 1$ **to** q **do**
    $\varphi_i \leftarrow (1-\tau) \times \frac{w_i}{\sum_{j=1}^{k} w_j} + \tau \times \frac{1}{q}$
return sample from $\vec{\theta}$ according to $\vec{\varphi}$ distribution

Real-World Feedback

Once a strategy has been chosen and a seed set identified:

- we test S in the real-world (posting on Twitter, flyers in a city,…)
- in round n, we get activation feedback composed of activated nodes A_n , and feedback set F_n — tuples (i, j, a_{ij}) for every affected edge

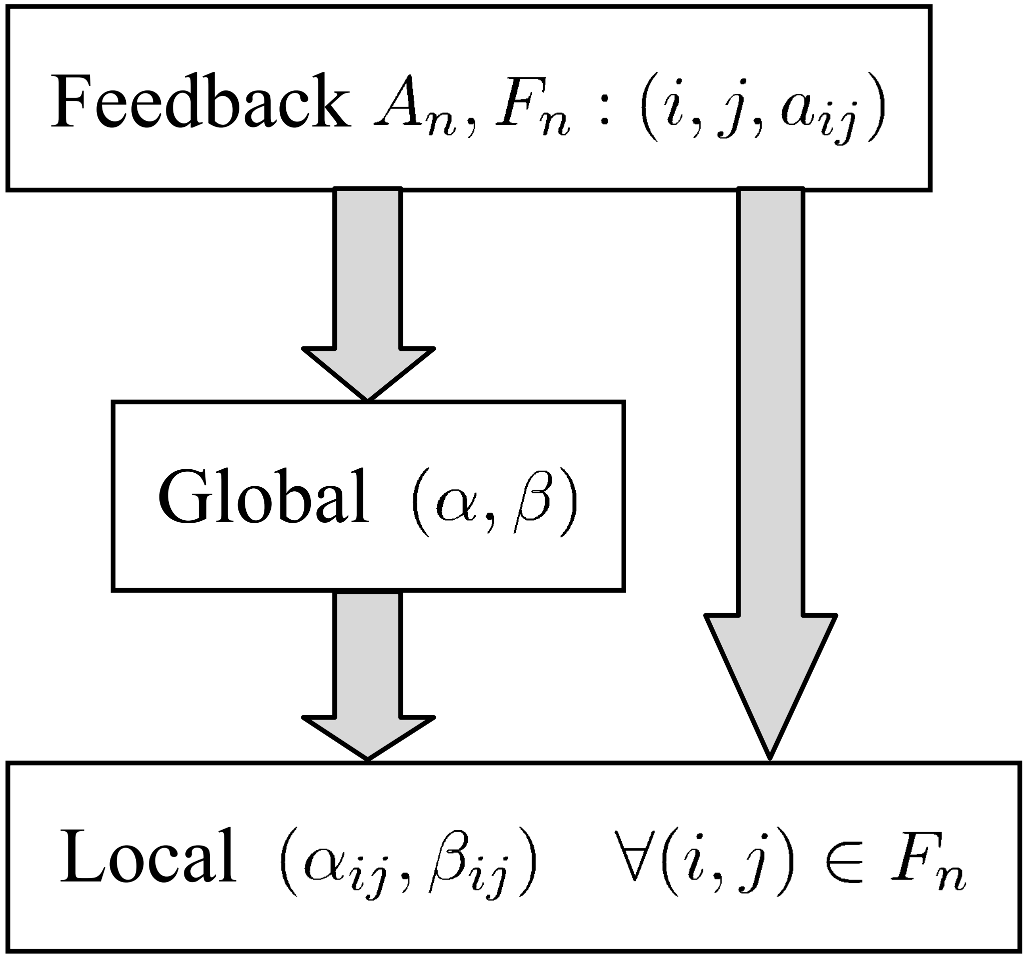
`Update` Step

Two approaches to **`Update`**:

- local update: each edge in the feedback is updated in a Bayesian manner

- global update: each edge in the graph is updated using methods such as maximum likelihood or least squares regression

- can also be combined



Local Update

Beta distribution is a conjugate prior of the Bernoulli distribution — the update is straightforward:

- success $a_{ij} = 1 \implies P_{ij} \sim \text{Beta}(\alpha_{ij} + 1, \beta_{ij})$
- failure $a_{ij} = 0 \implies P_{ij} \sim \text{Beta}(\alpha_{ij}, \beta_{ij} + 1)$
- same as counting the number of successful and failed activations for each edge

Global Update

Only using local update might be too sparse — especially for low influence probabilities, can lead to over reliance on the prior.

Solution: update also the prior for all edges, using all the feedback history

Global Update

Ordinary Least Squares (LSE): update via least squares estimation, from the formula of a spread of a node:

$$\sigma_n(\{s\}) = 1 + \sum_{\substack{(s,i)\in E \\ i\notin \mathcal{A}_n}} p_{si} \times \sigma_n(\{i\}) + \sum_{\substack{(s,i)\in E \\ i\in \mathcal{A}_n}} p_{si} \times (\sigma_n(\{i\}) - 1)$$

which leads to

$$(|A_n| - 1)\beta = (1 - |A_n|)(t_s + 1) + (h_s + o_s)\hat{\sigma}_n - (h_{as} + a_s)$$

$$x_n\beta = y_n.$$

$$\hat{\beta} = (\vec{x} \cdot \vec{y}) \ / \ (\vec{x} \cdot \vec{x})$$

Global Update

Maximum Likelihood (MLE): assume edges are independent:

$$\mathcal{L}(F_n \mid \alpha, \beta) = \prod_{(i,j,a_{ij}) \in F_n} \frac{(\alpha + h_{ij})^{a_{ij}} (\beta + m_{ij})^{1-a_{ij}}}{\alpha + \beta + h_{ij} + m_{ij}}.$$

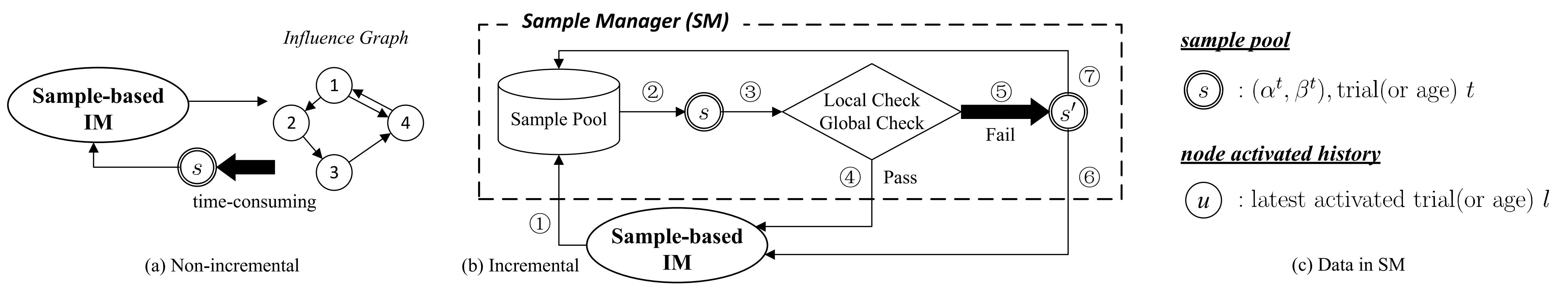
and the parameters can be estimated from

$$\sum_{(i,j,a_{ij}) \in F_n, a_{ij}=1} \frac{1}{\alpha + h_{ij}} = \sum_{(i,j,a_{ij}) \in F_n, a_{ij}=0} \frac{1}{\beta + m_{ij}}$$

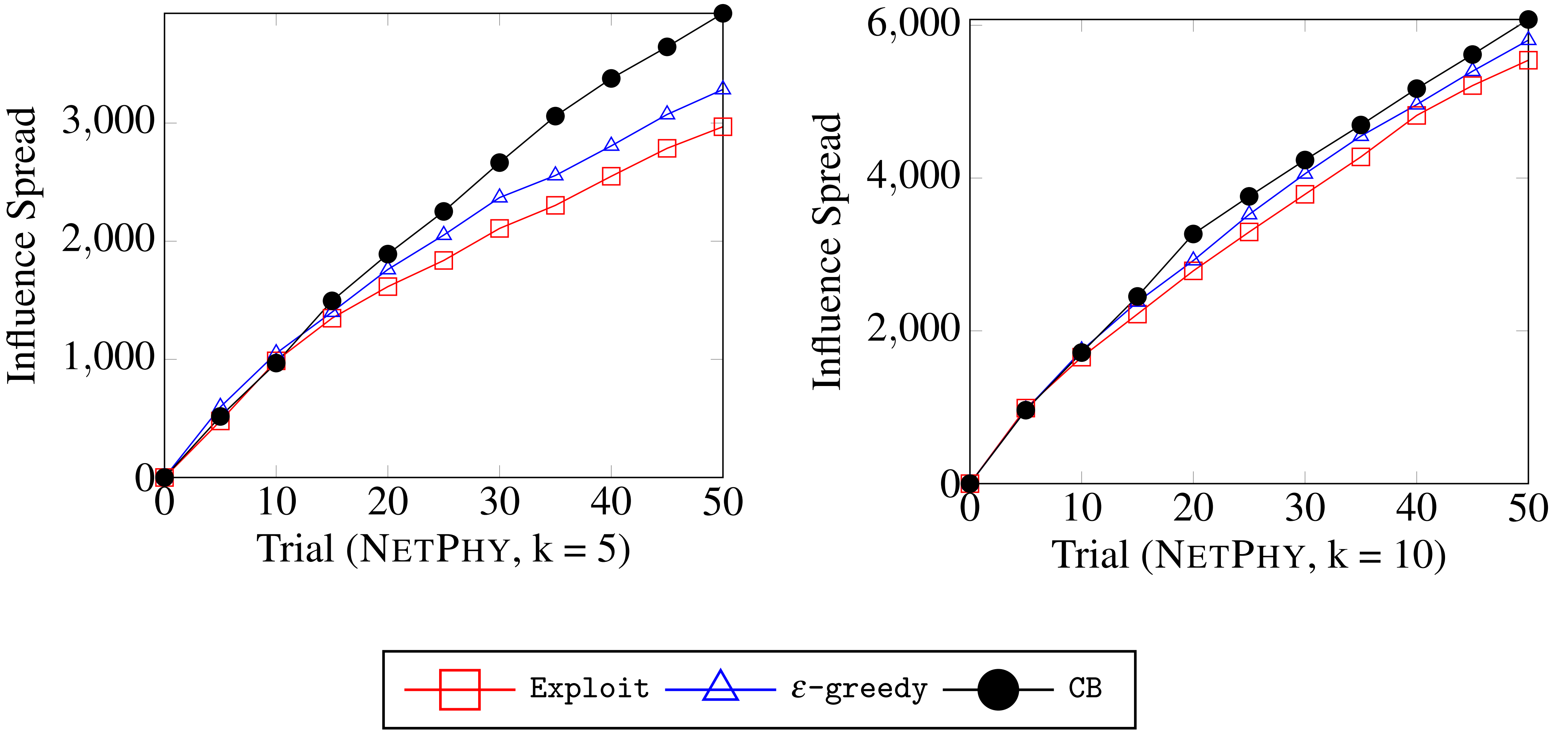
Sampling Optimization

Even advanced algorithms rely on sampling for influence estimation — costly over multiple rounds

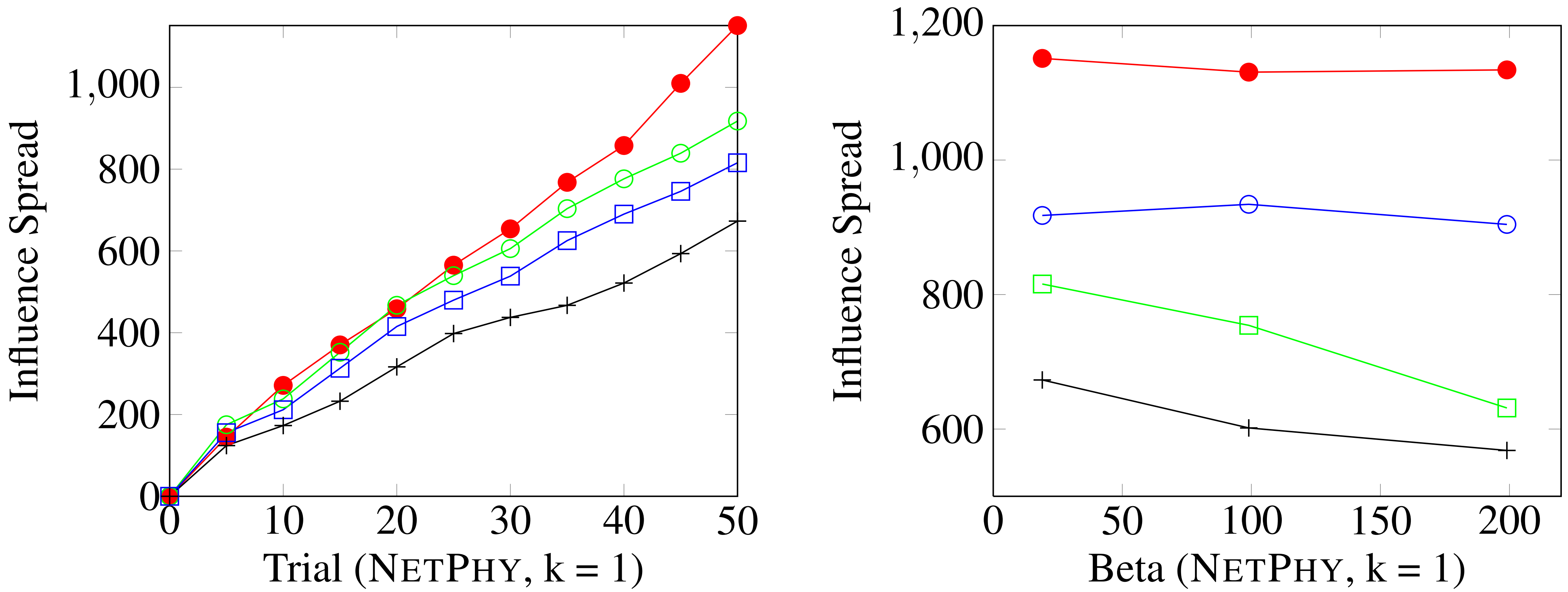
- incremental optimization approach — reuse of samples between rounds in little-affected parts of the graph



Results: effectiveness of explore-exploit strategies

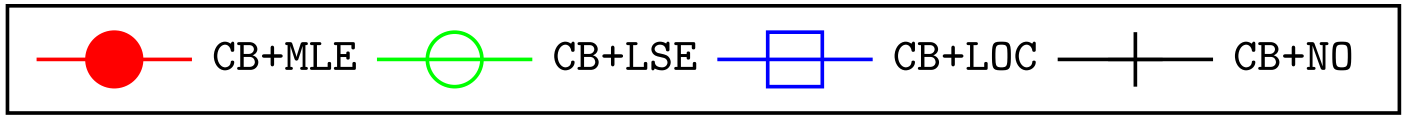


Results: effectiveness of update methods

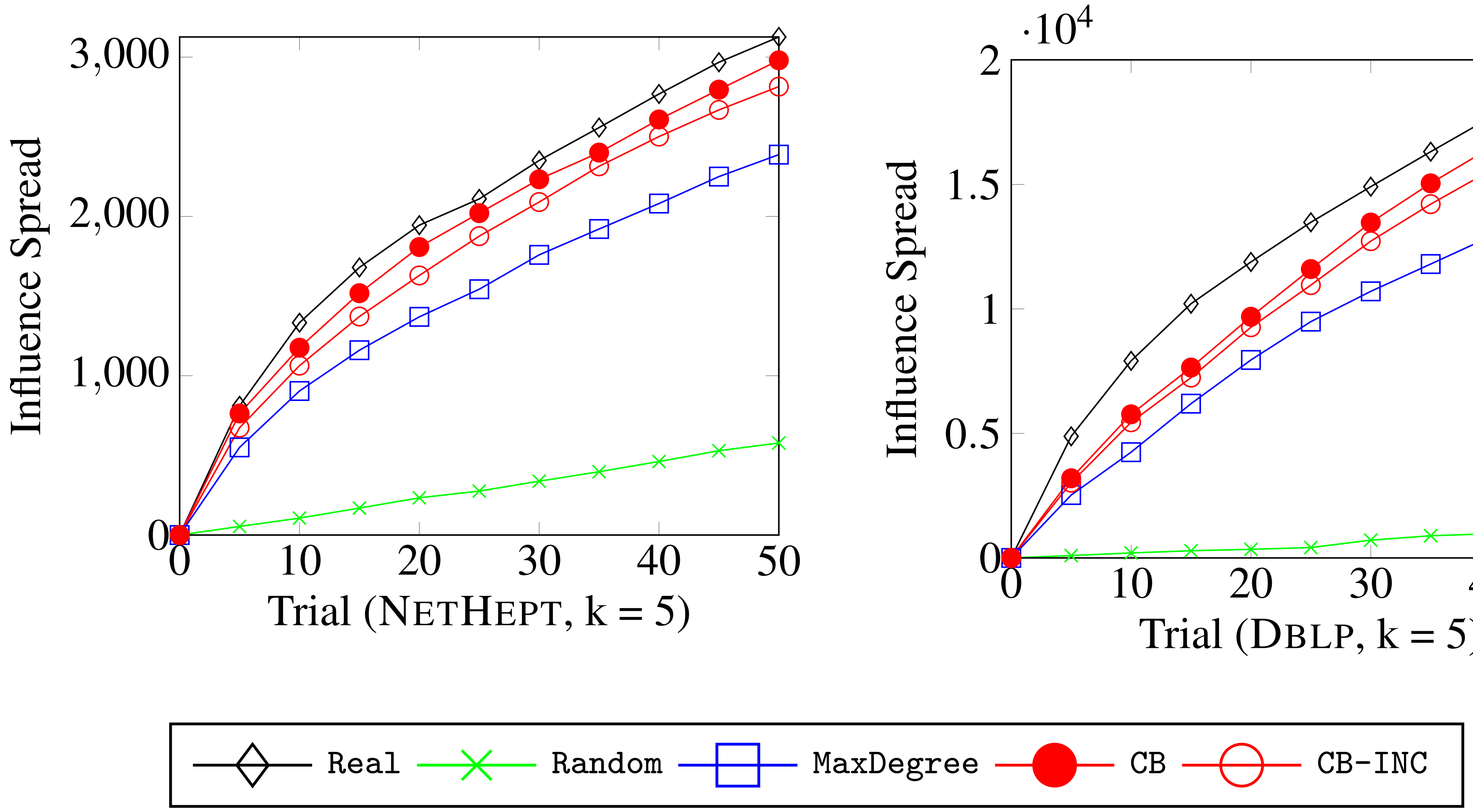


(a) Different updates

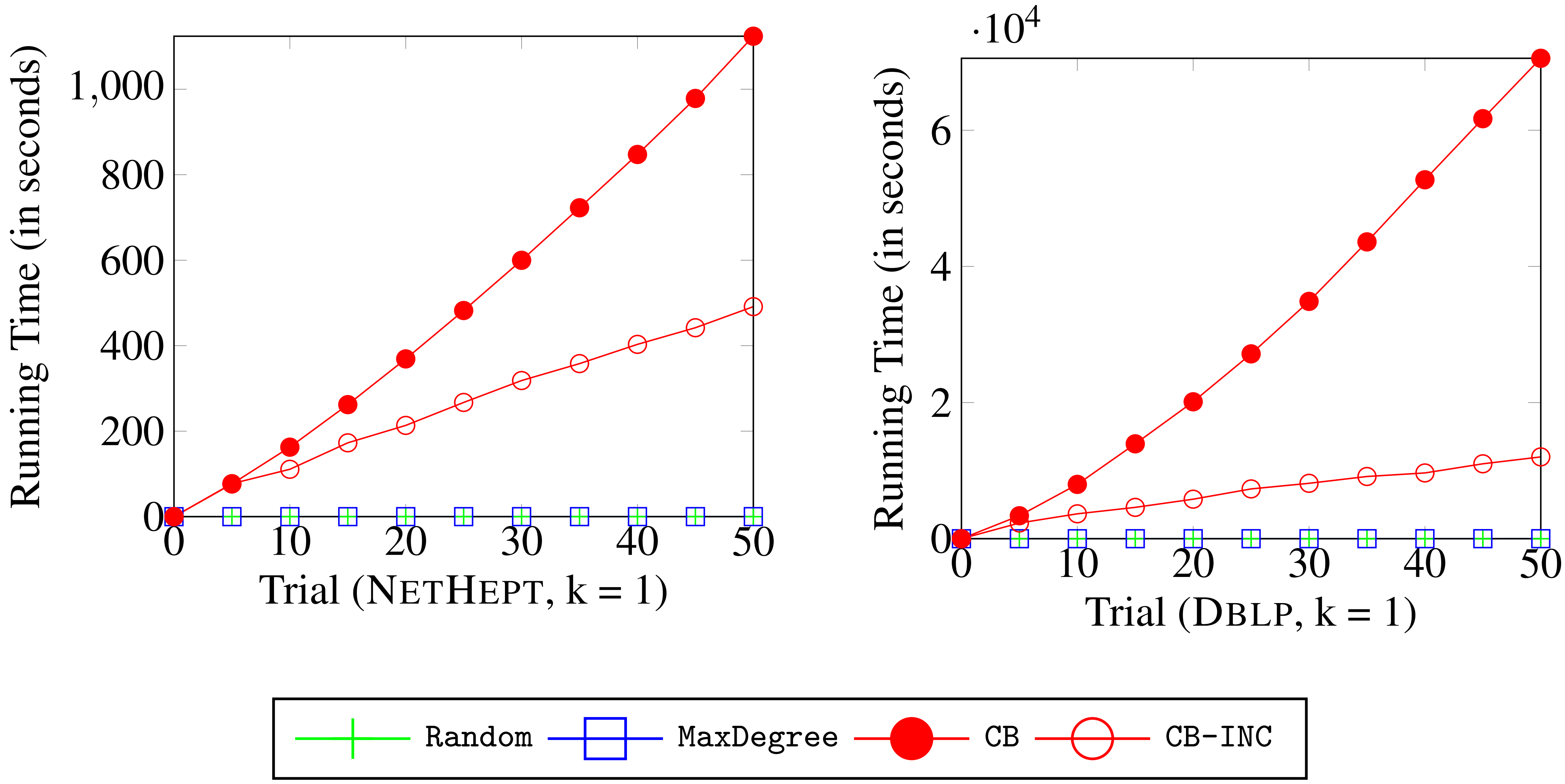
(b) Effect of priors



Results: effectiveness versus heuristics



Results: efficiency of sample reuse



Research Perspectives

- scalability is still a big issue in influence maximisation — even more so in the online setting
- adapting the framework to other influence models (threshold, credit distribution)
- learning also the influence model — do not rely on "synthetic" models such as independent cascade and threshold