A Linear Time Active Learning Algorithm for Link Classification

Nicolò Cesa-Bianchi* Claudio Gentile° Fabio Vitale* Giovanni Zappella*

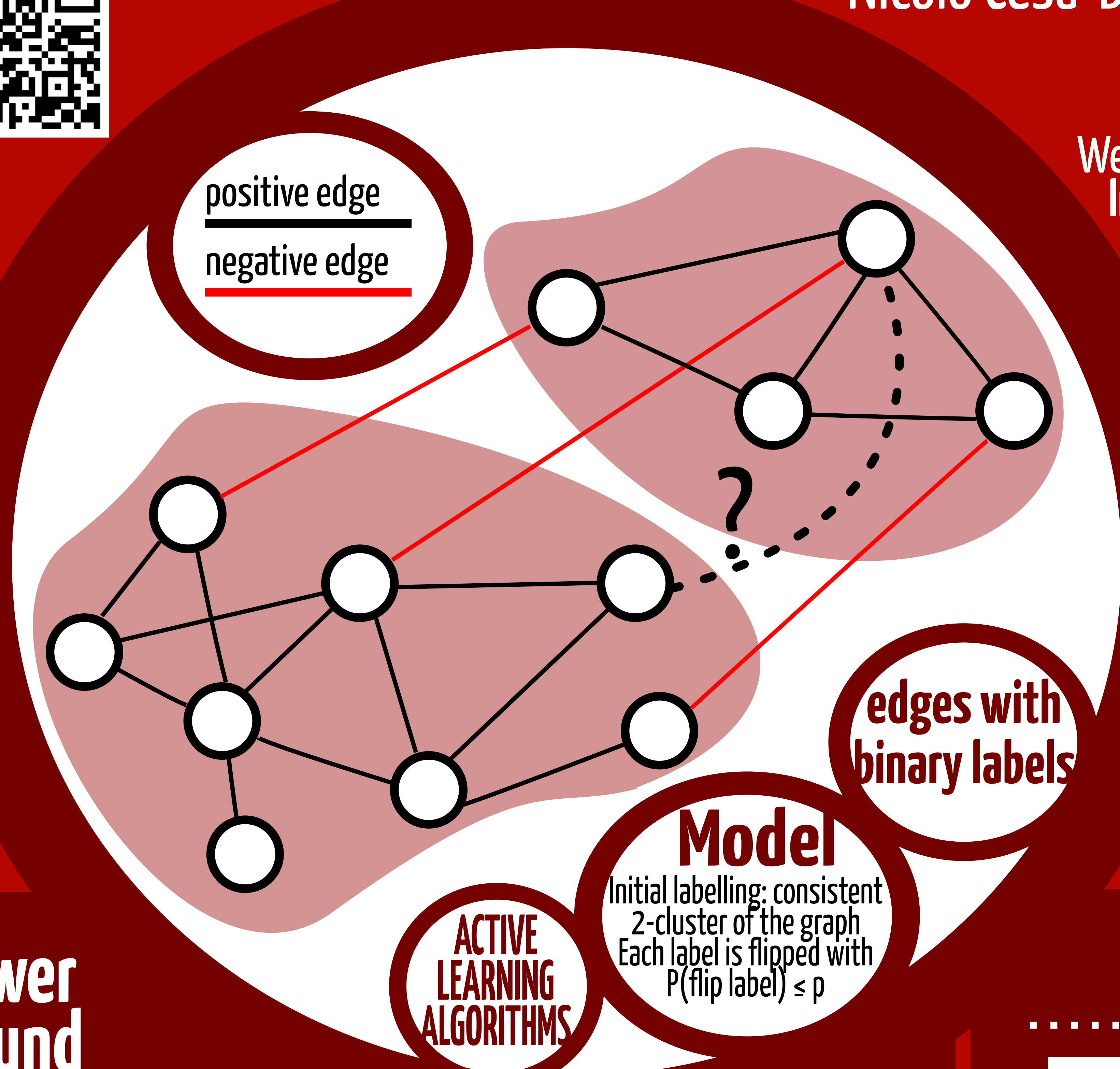
* → Università degli Studi di Milano °→Università dell'Insubria

W71

We present three efficient active learning algorithms for **link classification** in **signed networks**.
Signed networks are graphs whose edges carry a sign representing the **positive or negative nature** of the relation.
Applicative domains such **e-commerce, social networks** and **biology** offer examples of this kind of networks.

Most of the heuristics for link classification are summarized by the motto **"the enemy of my enemy is my friend"** (i.e. see [1,2,3]).

In this paper we assume a 2-cluster structure with a p-stochastic label assignment.
Each edge's label has P(sign_is_flipped) ≤ p, where the initial labelling is a consistent 2-cluster.



edges with binary labels

Model
Initial labelling: consistent 2-cluster of the graph
Each label is flipped with P(flip label) ≤ p

ACTIVE LEARNING ALGORITHMS

Lower Bound

Given a training set Eo

M ≥ p | E \ Eo|

for any undirected graph

Breadth-first Spanning tree

Using a simple BF-spanning tree it is possible to bound the expected number of mistakes as

E[M] ≤ 2 Diam p | E \ Eo |

StarMaker

For any undirected graph with $|E| = \Omega(|V|^{3/2})$ the expected number of mistakes made by **StarMaker** is

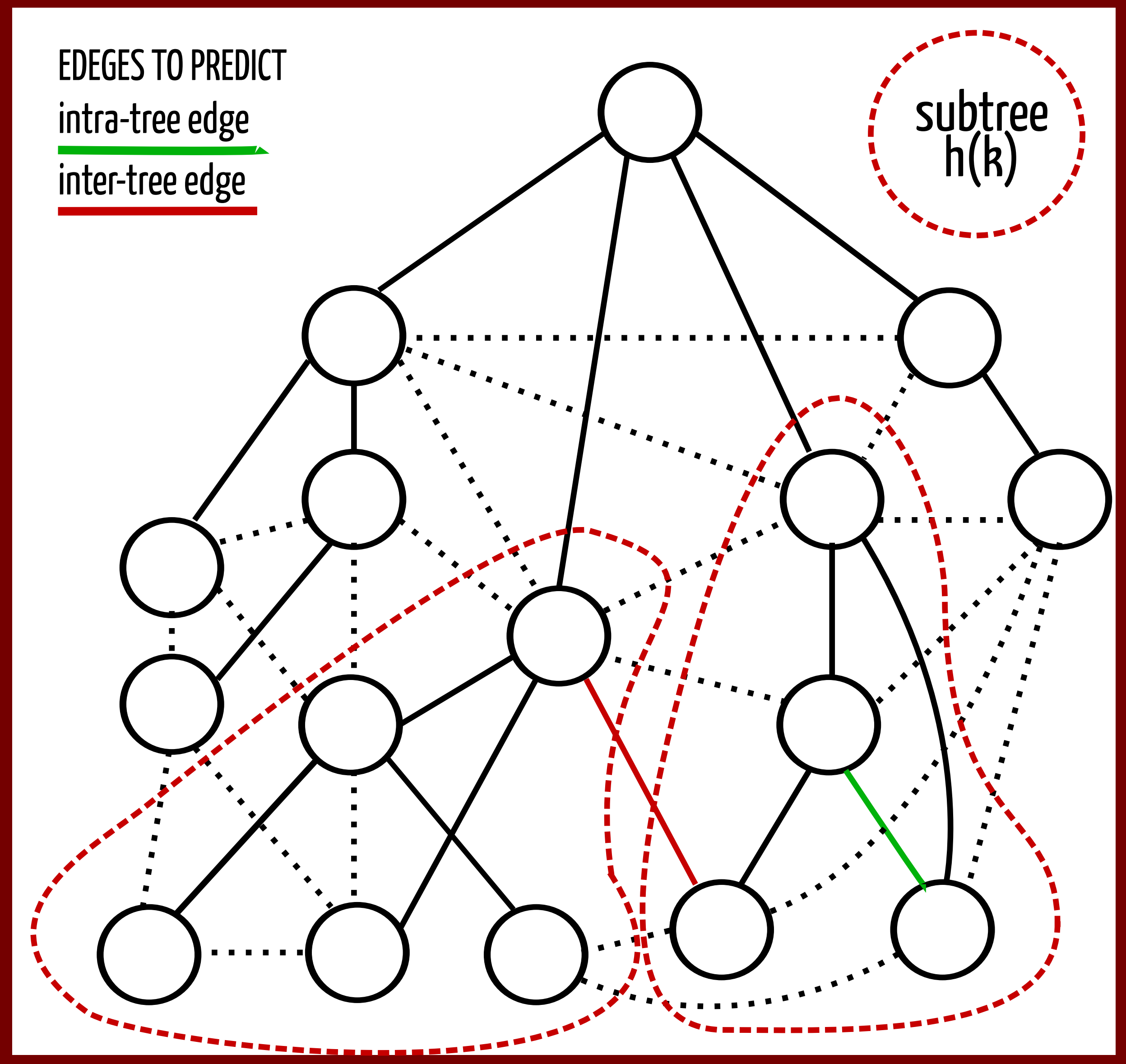
E[M] ≤ 5 p |E \ Eo|

while the query set size is upper bounded by $O(|V|)^{3/2}$.

It is possible to reduce the query set size by a factor of $k^{3/2}$ using **TreeLetStar(k)** with a mistake bound

E[M] = O(min{k,Diam}) p |E \ Eo|

TreeCutter



analysis is available in the paper

Running times

When the query set size is not larger than the test set size, the running times for our algorithms are:

O(|E|) for **TreeCutter(k)**
O(|E| + |V| log |V|) for **StarMaker**
O(|E| + |V|/k log |V|/k) for **TreeLetStar(k)**

We present 3 algorithms with different tradeoffs between mistakes and query set size

Experiments

Setup

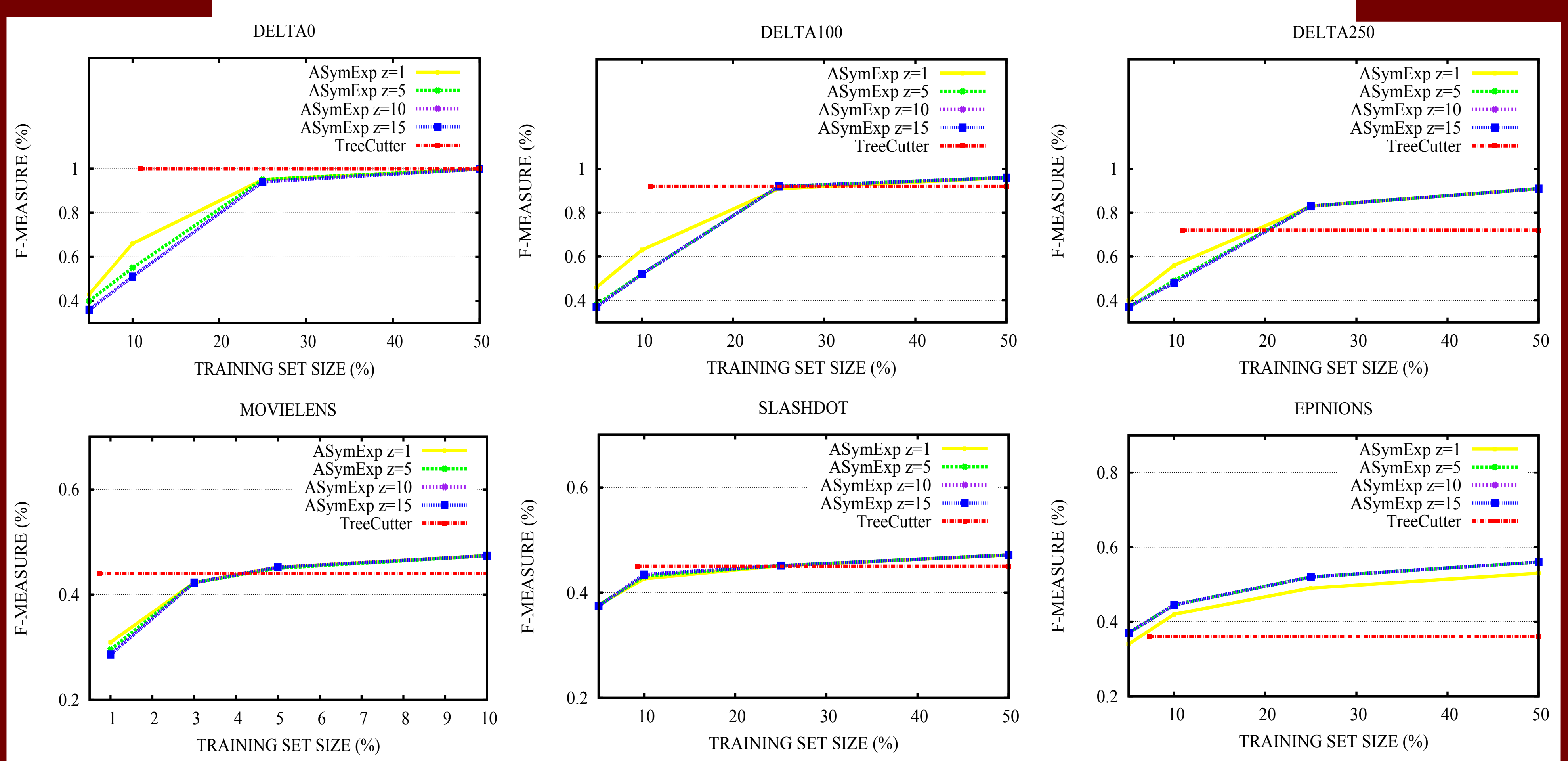
We used different training set sizes for ASymExp (indicated on the x-axis) and a fixed amount (|V|-1 edges) for TreeCutter. Since classes are very unbalanced, we report the F-Measure averaged on 10 runs for each combination of parameters.

Algorithms

We compared our simplest algorithm (**TreeCutter**) to the heuristics presented in [1], but we reported only the best of those algorithms (**ASymExp**) with different settings of the parameter

Datasets

DELTA* are synthetic datasets created in order to control the number of flipped labels. (e.g. DELTA100 means 100 labels have been flipped in the original consistent 2-cluster).
MovieLens is a graph of users created from movies (normalized) co-rating.
SlashDot and **Epinions** are real-world social networks with natively signed edges.



These results are obtained with the simplest of our algorithms that requires only |V|-1 edges (a spanning tree of the graph) as query set.

Our algorithm is **faster** than its competitors, is **extremely simple to implement** and in most of the cases outperforms its competitors using just a fraction of their training set.

Short Bibliography

[1] Kunegis et al. "The Slashdot Zoo: Mining a social network with negative edges" -- WWW 2009
[2] Cesa-Bianchi et al. "A correlation clustering approach to link classification in signed networks" -- COLT 2012
[3] Leskovec et al. "Predicting positive and negative links in online social networks" -- WWW 2010