



**HAL**  
open science

## Seeded Influencer Ranking

Marco Bressan, Philippe Caillou, Cyril Furtlehner, Michèle Sebag, Fabien Barzic

► **To cite this version:**

Marco Bressan, Philippe Caillou, Cyril Furtlehner, Michèle Sebag, Fabien Barzic. Seeded Influencer Ranking. Big Data Business Convention, Nov 2015, Jouy-en-Josas, France. hal-03832447

**HAL Id: hal-03832447**

**<https://hal.science/hal-03832447v1>**

Submitted on 27 Oct 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Seeded Influencer Ranking

Marco Bressan and Philippe Caillou and Cyril Furtlehner and Michèle Sebag and Fabien Barzic



## Goal

INPUT: Given a social network (graph structure and message contents), expert labels a handful of persons (social nodes) as influencers  
 OUTPUT: Learn

### An influence score

and use it to retrieve other influencers.

## The SIR Algorithm

**Require:** Social network: set  $S$  of persons, tweets (graph and contents), blogs, articles...

**Require:** Seed Influencers:  $I \subset S$

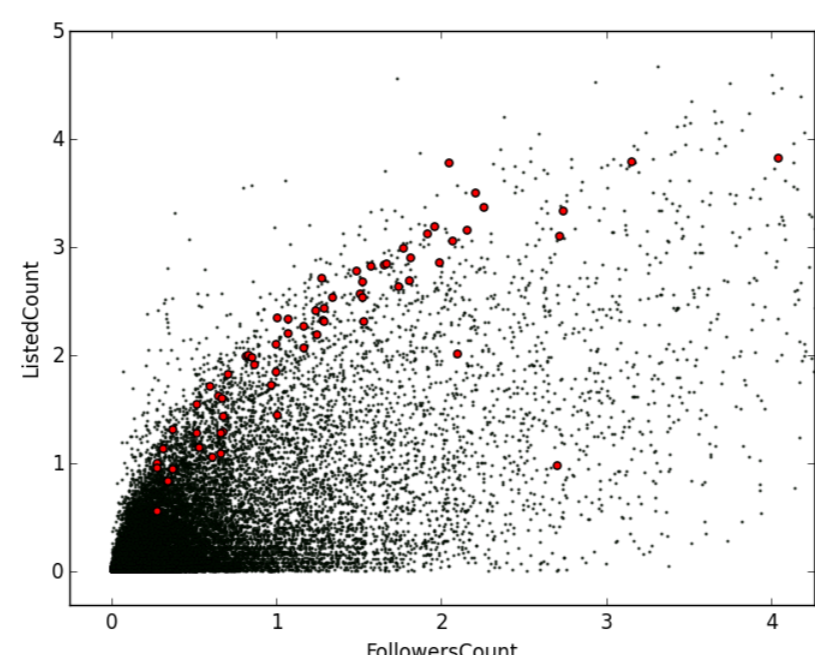
- 1: Build representation  $\mathbb{R}^d$ : words of  $I$ -people, graph-based features, social-based features
- 2: Map each person  $x$  onto  $\mathbb{R}^d$
- 3: Find  $h: \mathbb{R}^d \mapsto \mathbb{R}$  s.t.

$$h(x, x \in I) \gg h(x, x \notin I)$$

- 4: Use  $h$  as influence score
- 5: **Return** the top- $k$  individuals according to  $h$
- 6: Validation: rank of (hidden) test influencers

## Claim

- Influencer identification is a **supervised learning problem**
- from a social network, popularity and content-based features
- No syntactic influencer identification function



Twitter followers vs. ListedCount for 50k candidates - red circles for the true influencers; PR dataset

Influence, like beauty, lies in the eye of the beholder

## Empirical Validation, 6 domains

- *Fashion*: 18 influencers
- *High-Tech*: 69 influencers
- *Pub. Relations*: 39 influencers
- *Time*: 25 influencers
- *Wine*: 28 influencers
- *Human resources*: 99 influencers

EXPERT  
 professional fashionist  
 media specialist  
 PR manager  
 Time Magazine  
 Wikipedia  
 HR-focused blog.

	Application set	Fashion	High-Tech	PR	Times	Wine	HR
Train	influencers	13	52	29	19	21	74
	total	7,5M	7,5M	7,5M	7,5M	7,5M	7,5M
Test	influencers	5	17	10	6	7	25
	total	2,5M	2,5M	2,5M	2,5M	2,5M	2,5M

## Data

- 110M retweets (10% all retweets, Nov. 2014)
- 10.5 million nodes (persons)

## Representation

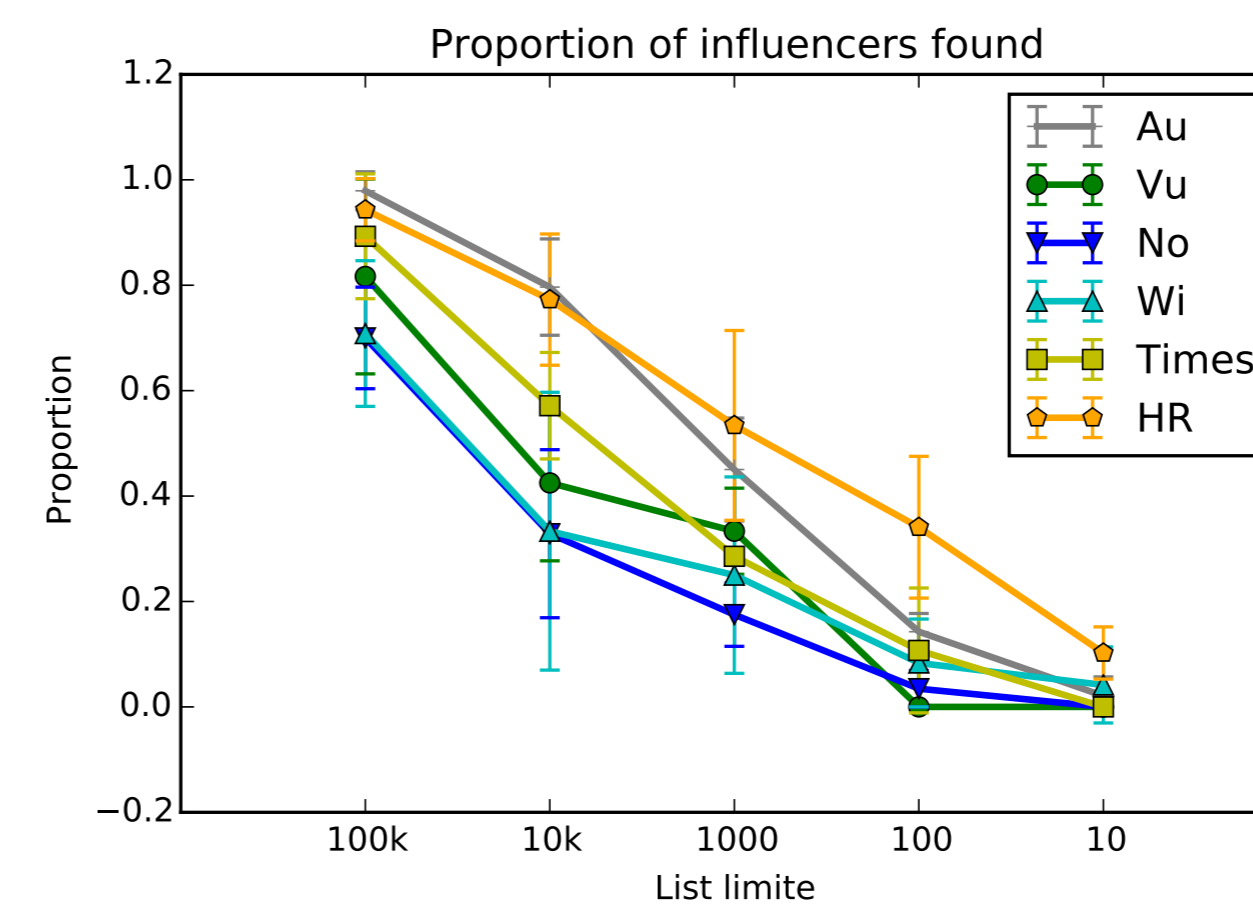
111 application-dependent features

- **Social features** (e.g. followers, favorites, apparition in lists, friends, status count)
- **Network features** (Pagerank score, inbound and outbound links, inbound/outbound ratio, harmonic and closeness centrality)
- **Content features** (Top 100 words from influencers) topic-dependent

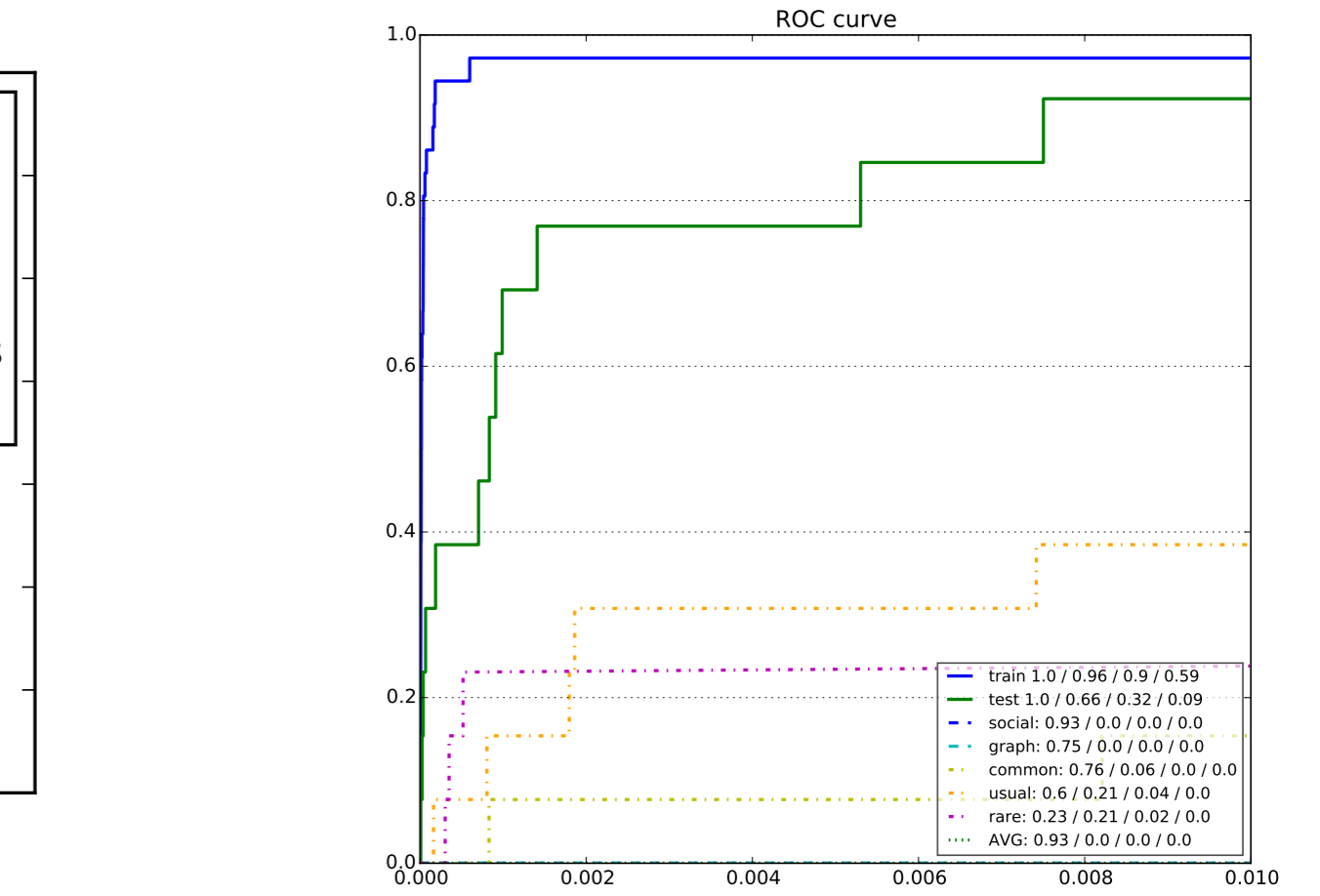
## Experimental setting

- 4-fold Cross Validation (train on 3/4, test on 1/4, average)
- Optimize Area Under the Curve AUC @ 1000  
 Find  $\text{argmin} \{ \mathcal{F}(h) = \sum \text{rang}(x_i), x_i \in S \cap \text{train}, \text{rang}(x_i) < 1,000 \}$
- Performance indicator: recall @k  $\sum \text{rang}(x_i), x_i \in S \cap \text{test}, \text{rang}(x_i) < k$

## Results

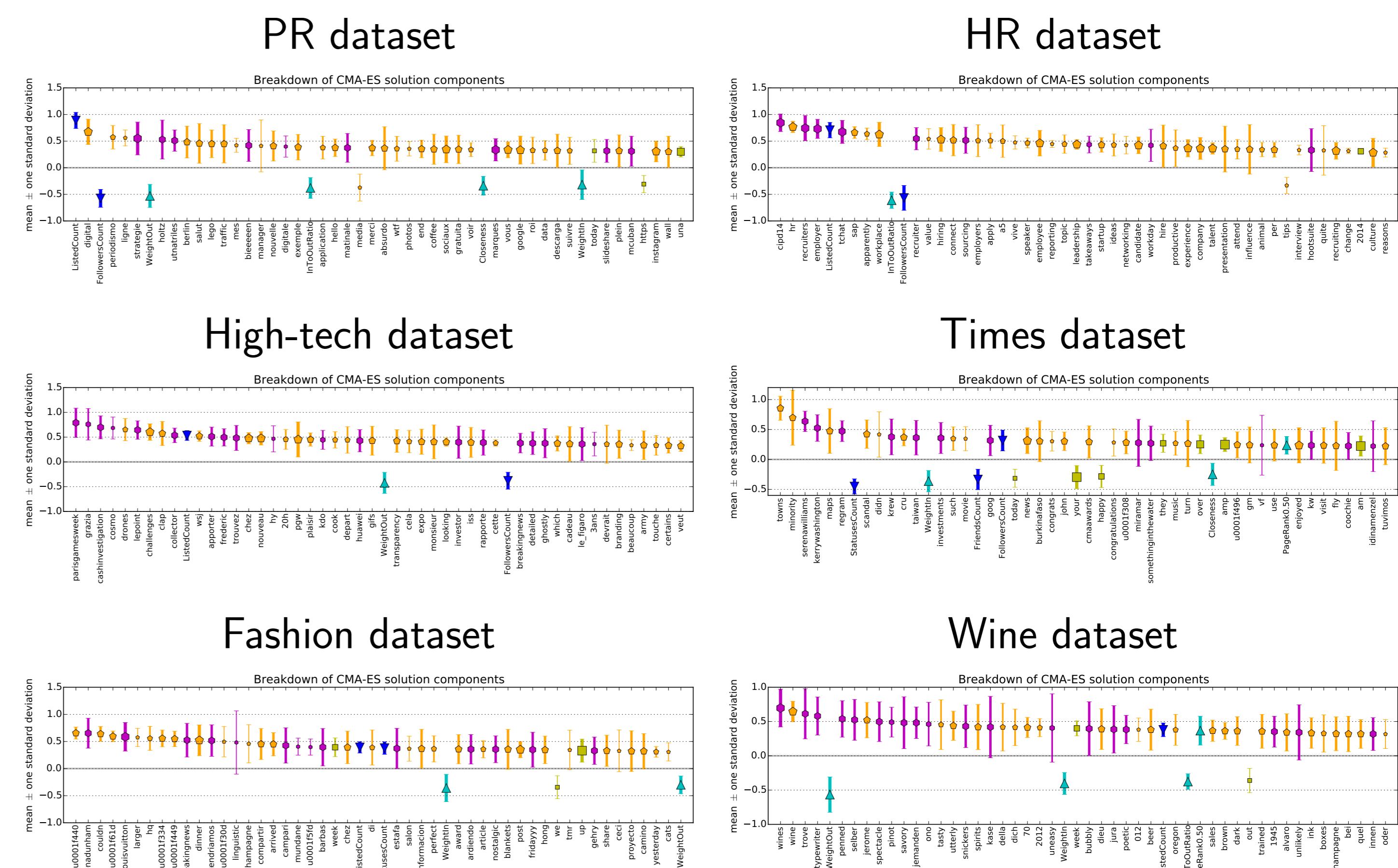


Recall for each dataset % influencers in top k



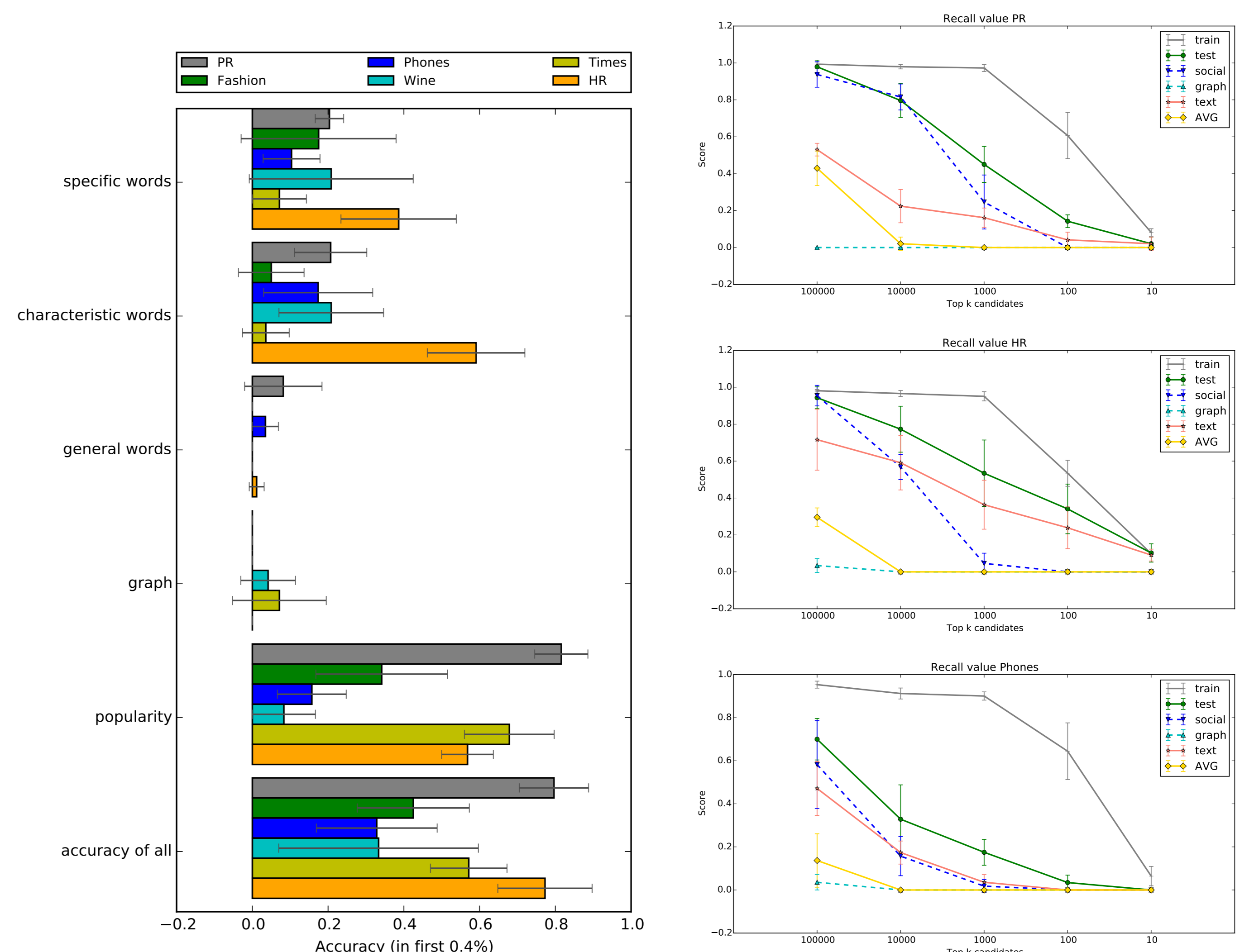
ROC curve example on PR dataset sensitivity wrt feature set

## Feature & words relevance wrt domain



Average learned weights (on all folds and runs) for the top 30 features for each dataset. The width of the bar represents the number of folds where the feature exists (content-dependent features vary depending on the fold)

## Lesion studies: importance of feature categories



Recall for a single (all) feature category Recall @10 ... 10,000 for PR, HR and High-Tech domains  
 Recall sensitivity w.r.t. feature category

## Discussion

- Supervised influencer identification is successfully validated
- Model stability over folds for a single domain
- Different models for different domains: **no syntactic influence score.**

Contact: [caillou@lri.fr](mailto:caillou@lri.fr)