Seeded Influencer Ranking

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

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: $I \subset S$

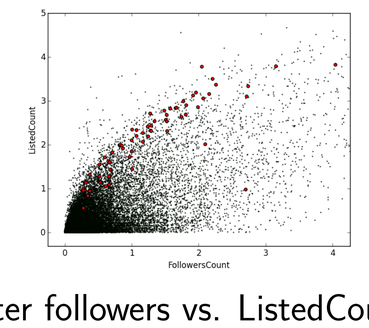
1: Build representation $\mathbb{R}^d$: words of I-people, graph-based features, social-based features
2: Map each person x onto $\mathbb{R}^d$
3: Find $h : \mathbb{R}^d \mapsto \mathbb{R}$ s.t.

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

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 feeatures
- 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

	EXPERT
• *Fashion*: 18 influencers	professional fashionist
• *High-Tech*: 69 influencers	media specialist
• *Pub. Relations*: 39 influencers	PR manager
• *Time*: 25 influencers	Time Magazine
• *Wine*: 28 influencers	Wikipedia
• *Human resources*: 99 influencers	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

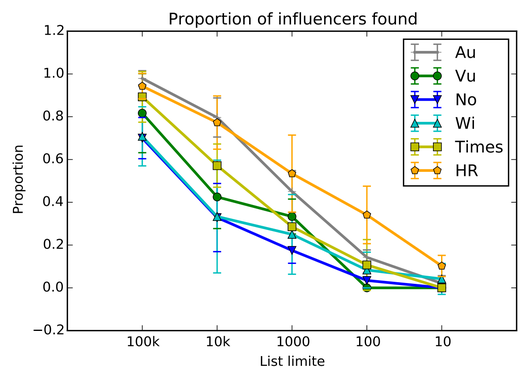
111 application-dependent features

- **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

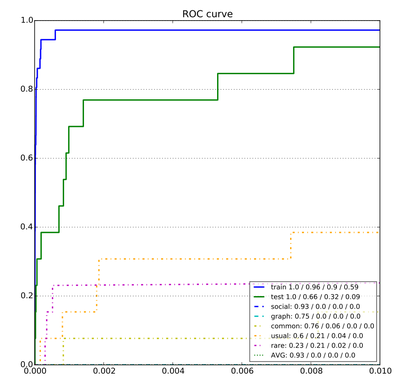
Experimental setting

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

Results

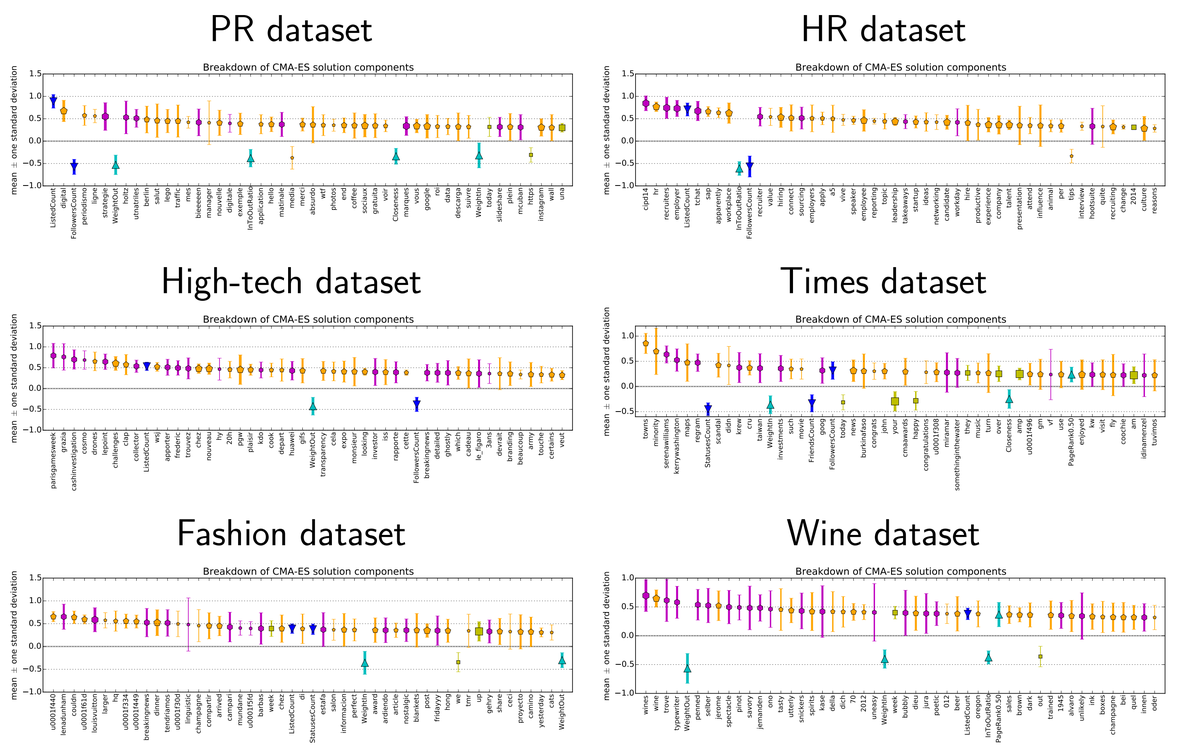


Recall for each dataset
% influencers in top k



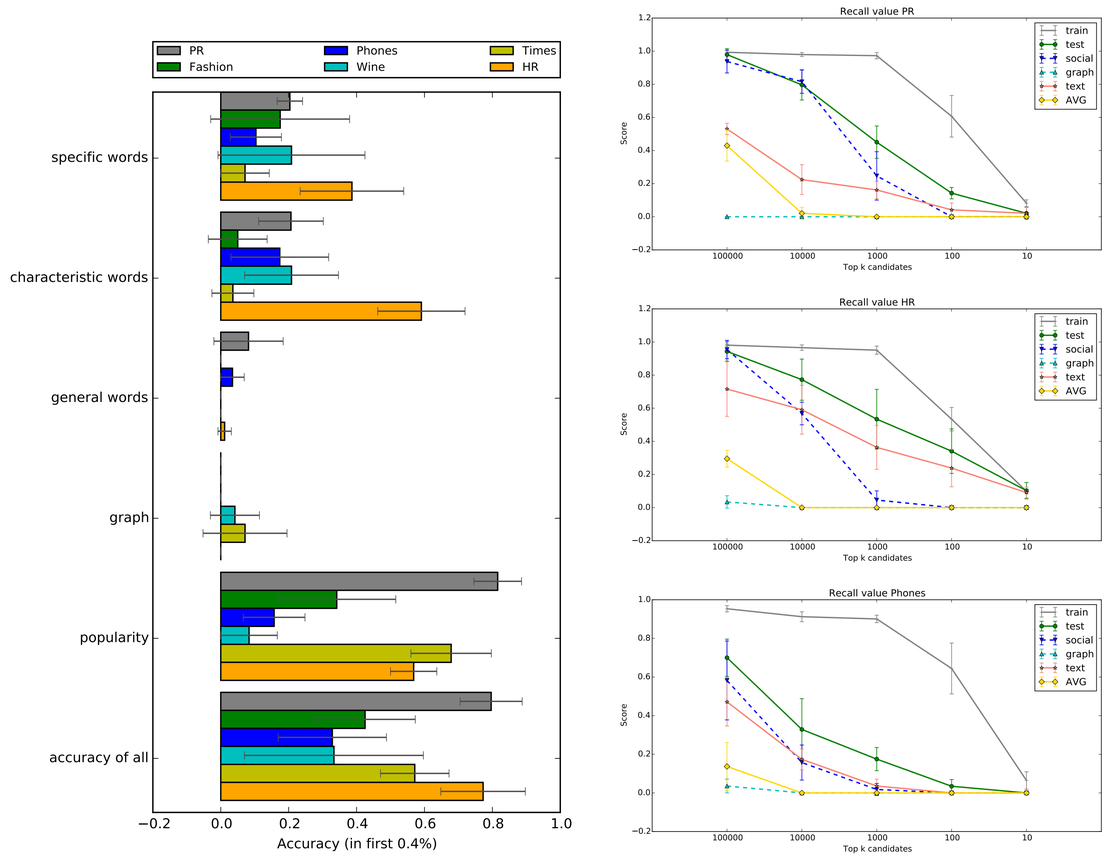
ROC curve example on PR dataset
sensitivity wrt feature set

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 sensitity 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**