Learning Distributed Representations of Users for Source Detection in Online Social Networks

The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2016

Simon Bourigault, Sylvain Lamprier and Patrick Gallinari

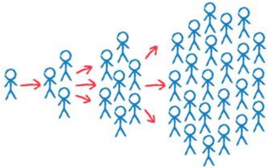
Contents

Background

Field

Facebook

- Social Networks
- Network Diffusion
- Source Detection



https://www.linkedin.com/pulse/2014091814
7859692-social-network-301-what-is-virality

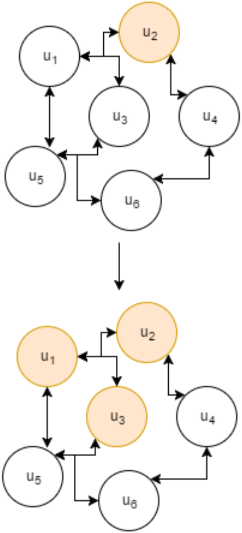
Contributions

- Introducing representation learning approach to the field of source detection delivering a robost model that handles data sparsity well
- Does not require the influence graph to be known
- Tested on real life data and surpassing other baseline approaches
- Provides an extension that further improves the results

Related Work

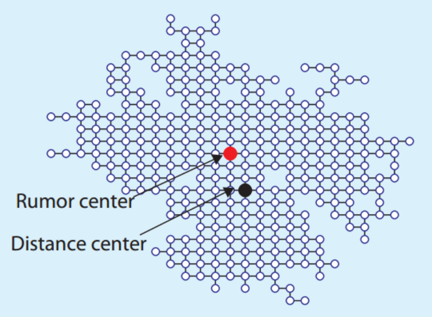
Diffusion Prediction

- Susceptible-Infected framework
  - Varies in how to reverse the process of diffusion to predict the source



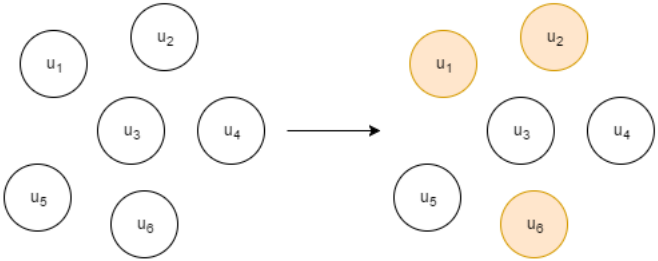
Rumor centrality

- Detecting sources of computer viruses in networks: theory and experiment
  - Shah, D. and Zaman, T.



Theory

Diffusion Episodes



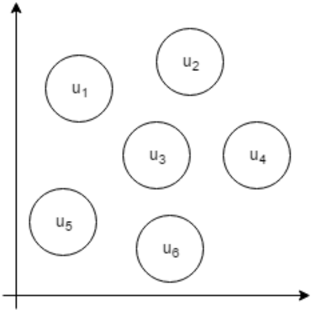
Group of users before and after a diffusion episode

Diffusion Episode Definition

$$D = \{(u_i, t_i), (u_j, t_j)...\}$$

Representation Learning Model

- Latent space
  - Latent
  - Dimensionality Reduction
  - Projection onto euclidean space with *d* dimensions
  - Distance correlates with chance of being the source
- Receiver and Sender embeddings



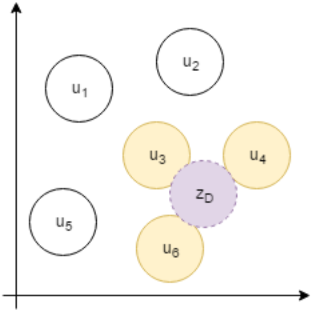
Source Prediction Model

Averaged Representation of Infected Users

$$z_D = \phi(\hat{U}_D) = \frac{1}{\hat{U}_D} \sum_{u_i \in \hat{U}_D} z_i$$

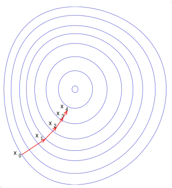
Diffusion Source

$$s^* = argmin_{u_i \in U / \hat{U}_D}(||w_i - z_D||)$$



Learning Step

- Stochastic Gradient Descent



Update Step

$$L(\Omega, Z) = \sum_{D \in \mathscr{D}} \sum_{u_i \notin U_D} h(||w_i - z_D||^2 - ||w_{s_D} - z_D||^2)$$

Regularization

$$L(\Omega, Z) + \lambda \sum_{u_i} ||w_i - z_i||^2$$

Learning Step

Algorithm 1. Representation Learning for Source Detection

Data:
$\mathcal{U}$: Users set ;
$\mathcal{D}$: Learning set of diffusion episodes ;
d : Number of dimensions
ϵ : Gradient step size;
Result:
$Z = \{\forall u_i \in \mathcal{U} : z_i \in \mathbb{R}^d\}$; $\Omega = \{\forall u_i \in \mathcal{U} : \omega_i \in \mathbb{R}^d\}$;

foreach $u_i \in \mathcal{U}$ **do**
  initialize z_i with random value in $[-1, 1]^d$
  initialize ω_i with random value in $[-1, 1]^d$
end
while *non-convergence* **do**
  Draw an episode $D \in \mathcal{D}$;
  Draw $u_j \notin \mathcal{U}_D$;
  Compute z_D with formula 1 ;
  $d_s \leftarrow ||\omega_{s_D} - z_D||^2$;
  $d_j \leftarrow ||\omega_j - z_D||^2$;
  if $d_j - d_s < 1$ **then**
    $\omega_{s_D} \leftarrow \omega_{s_D} - \epsilon \times 2\,(\omega_{s_D} - z_D)$;
    $\omega_j \leftarrow \omega_j + \epsilon \times 2\,(\omega_j - z_D)$;
    forall $u_x \in \hat{\mathcal{U}}_D$ **do**
      $z_x \leftarrow z_x - \epsilon \times \frac{2}{|\hat{\mathcal{U}}_D|}\,(\omega_j - \omega_{s_D})$
    end
  end
end

Extensions

Inclusion of User Importance

$$z_D = \sum_{u_i \in \hat{U}_D} \frac{e^{\alpha_i}}{\sum_{u_j \in \hat{U}_D} e^{\alpha_j}} z_i$$

Integration of Content

$$z_D = \frac{1}{|\hat{U}_D|} \sum_{u_i \in \hat{U}_D} z_i + < w_D, \theta >$$

Evaluation

Datasets

	Users	Links	Episodes	Density
Artificial	100	262	10 000	2 %
Lastfm	1984	235 011	331 829	5 %
Weibo	5000	20 784	44 345	0.08 %
Twitter	4107	128 855	16 824	1 %

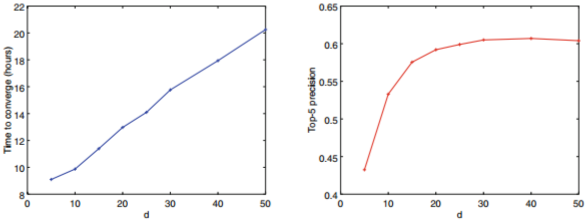
Dataset Statistics

Baseline Approaches

- OutDeg
  - Ranks sources by their out-degree
- Jordan Center
  - Predicted source is the node with the minimum longest distance to any infected
- Pinto's
  - Assumes infection delays follows a Gaussian law

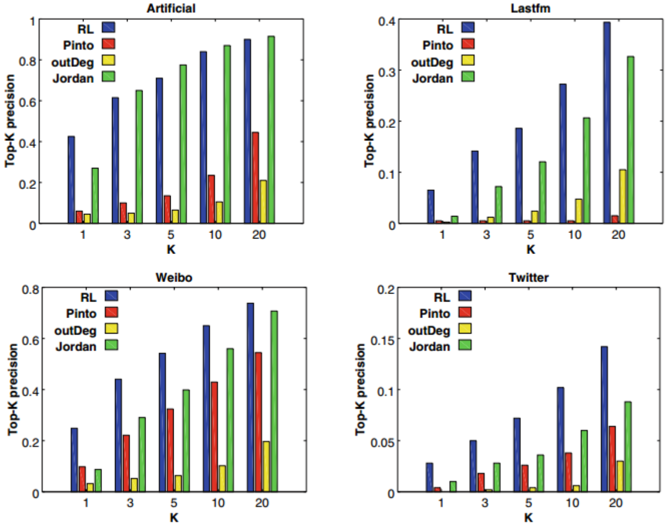
Training the Model

- Value of $d = 30$

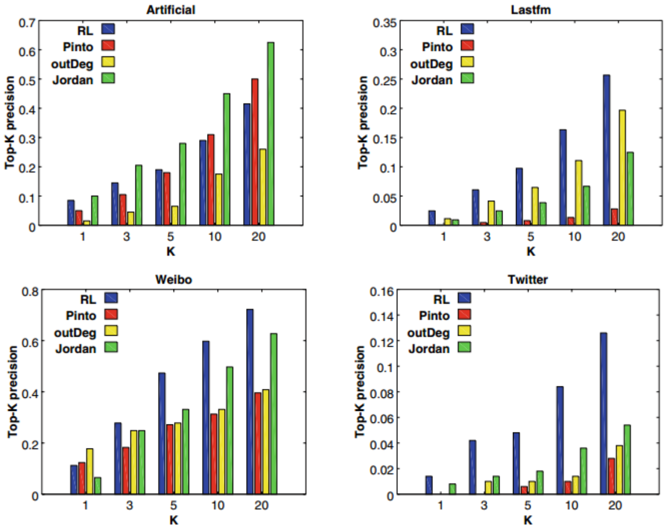


Convergence time and performance for various values of d on the Weibo dataset

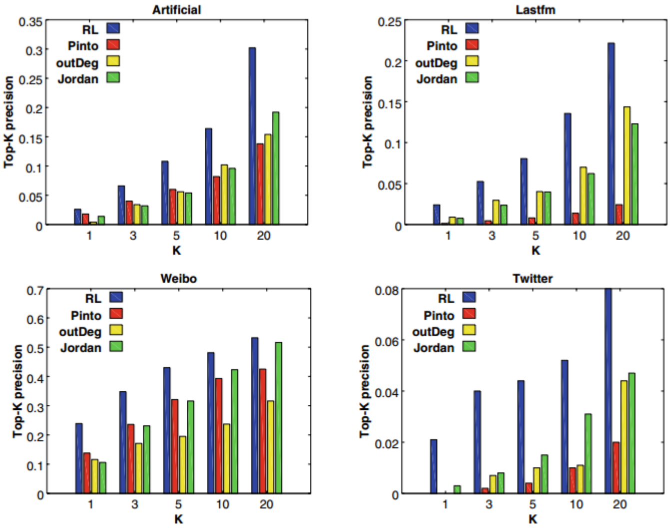
Results



Source detection with full cascades



Source detection with partial cascades (20%)



Source detection with partially observed cascades (20%)

Top-K	1	3	5	10	20
Twitter					
RL	0.020	0.042	0.058	0.099	0.141
RL w/weights	0.021	0.047	0.073	0.107	0.154
Gain	3 %	10 %	25 %	8 %	9 %
Lastfm					
RL	0.052	0.12	0.166	0.2545	0.374
RL w/weights	0.065	0.1335	0.175	0.2605	0.378
Gain	25 %	11 %	5 %	2 %	1 %
Weibo					
RL	0.31	0.51	0.59	0.72	0.82
RL w/weights	0.31	0.50	0.60	0.75	0.84
Gain	0 %	−2.3 %	+0 %	+4 %	+1 %

Source detection with user weights

Top-K	1	3	5	10	20
RL	0.028	0.05	0.072	0.102	0.142
RL w/content	0.043	0.069	0.099	0.128	0.179
Gain	56 %	38 %	38 %	26 %	26 %

Source Detection with Content Integration on the Twitter dataset

Criticism

Criticism

- Inconsistent reference strategy. Did not always refer to their figures and formulas
- Latent Space
  - Missing flow between concept and construction of representation model
  - No reference to the picture
- Size of dataset
- High concept, low technical
- Results of Content Integration for Twitter dataset