Early Identification of Pathogenic Social Media Accounts

Hamidreza Alvari, Elham Shaabani, Paulo Shakarian
Arizona State University
{halvari,eshaaban,shak}@asu.edu

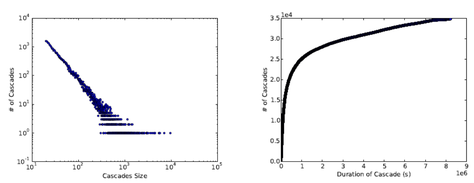
Abstract

PSM accounts such as terrorist supporters exploit communities of supporters for attacking social media. Early detection of PSMs is crucial as they are high likely to be *key* users in making a harmful message "viral". In this paper, we propose a time-decay causality metric to identify PSMs within a short time frame around their activity. We then propose a causal community detection-based classification algorithm to enhance the performance. The method is applied to groups of accounts sharing similar causality features and is followed by a classification algorithm to classify them as PSM or not. Unlike existing techniques we do not have to collect information such as network, cascade path, or content, as we only rely on action log of users.

Dataset

- 53M ISIS related tweets/retweets in Arabic (Feb, 22 2016 – May 27, 2016) using 290 Hashtags (e.g., #Terrorism, #BurqaState, #Rebels)
- Cascade sizes vary between 20 to 9,751 and take from 10 secs to 95 days (log-log plot shown below).
- We use no more than 50% of the dataset to ensure timeliness of the approach

Name	Value	
# of Cascades	35K	
# of Viral Cascades	6,602	
# of Tweets/Retweets	2,808,878	
# of Users	Suspended	Active
	64,484	536,609



Causal inference

- Kleinberg-Mishra Causality [Kleinberg-Mishra 2009]: measures how causal user '*i*' is in making a message 'viral'.

$$\epsilon_{K\&M}(i) = \frac{\sum_{j \in \mathbf{R}(i)} (p_{i,j} - p_{\neg i,j})}{|\mathbf{R}(i)|}$$

- Related users to user '*i*':

$$\mathbf{R}(i) = \{j | j \neq i \text{ and } i, j \text{ are } m\text{-related}\}$$

- Relative Likelihood Causality [Shaabani et. al., 2018]:

$$\mathcal{E}^I_{rel}(i) = \frac{\mathcal{S}(i,j)}{|\mathcal{R}(i)|} \qquad \mathcal{S}(i,j) = \begin{cases} \frac{\mathcal{P}_{i,j}}{\mathcal{P}_{\neg i,j}+\alpha} - 1, & \mathcal{P}_{i,j} > \mathcal{P}_{\neg i,j} \\ 1 - \frac{\mathcal{P}_{\neg i,j}}{\mathcal{P}_{i,j}}, & \mathcal{P}_{i,j} \leq \mathcal{P}_{\neg i,j} \end{cases}$$

- Neighborhood-based Causality [Shaabani et. al., 2018]:

$$\mathcal{E}^I_{nb}(j) = \frac{\sum_{i \in \mathcal{Q}(j)} \mathcal{E}^I_{K\&M}(i)}{|\mathcal{Q}(j)|} \qquad \mathcal{E}^I_{wnb}(j) = \frac{\sum_{i \in \mathcal{Q}(j)} w_i \times \mathcal{E}^I_{K\&M}(i)}{\sum_{i \in \mathcal{Q}(j)} w_i}$$

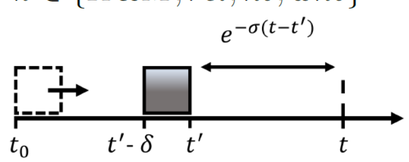
$$\mathcal{Q}(j) = \{i | j \in \mathcal{R}(i)\}$$

Decay-based Causality

- Causality of users may change over time
- Our score: (1) uses sliding time window, (2) deploys exponential decay function, (3) takes the average of the causality values computed over the windows

$$\xi^I_k(i) = \frac{1}{|\mathcal{T}|} \sum_{t' \in \mathcal{T}} e^{-\sigma(t-t')} \times \mathcal{E}^{\Delta}_k(i)$$

$$k \in \{K\&M, rel, nb, wnb\}$$



Visit us online at http://cysis.engineering.asu.edu

Approach

- We build graph **G** from action log **A** of users by connecting those who posted a message chronologically.
- We use Louvain algorithm to detect communities of users and use KNN to assign the label with the majority class

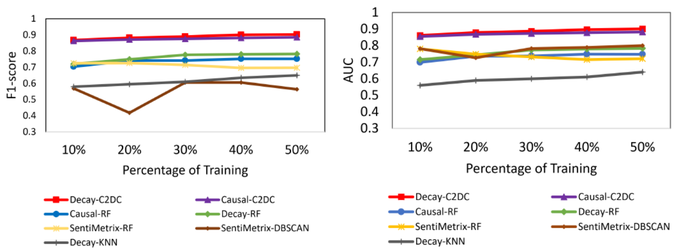
Algorithm 1 C^{2}DC

Require: Training samples $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$, tests $\{\mathbf{x}'_1, ..., \mathbf{x}'_n\}$, $\mathbf{G}$, k
Ensure: Predicted labels $\{y'_1, ..., y'_n\}$

```
C ← LOUVAIN(G)
for each x'_i do
    C_l ← C' ∈ C s.t. x'_i ∈ C'
    D ← {}
    for each x_j ∈ C_l do
        d_ij ← ||x'_i − x_j||_2
        D ← D ∪ {d_ij}
    end for
    K ← KNN(D, k)
    y'_i ← DOMINANT-LABEL(K)
end for
```

Results

- Community-based approach achieves the best performance



- We would like to see ***how many*** of PSMs who were active in first 10 days are correctly classified over time

Learner	True Positives/False Positives					Remaining
	02/22-03/02	03/02-03/12	03/12-03/22	03/22-03/31	03/31-04/09	
DECAY-C^{2}DC	3,072/131	286/0	0/0	0/0	0/0	0
CAUSAL-C^{2}DC	3,065/156	188/20	105/0	0/0	0/0	0
DECAY-KNN	2,198/459	427/234	315/78	109/19	96/0	213
DECAY-RF	2,472/307	643/263	143/121	72/68	28/0	0
CAUSAL-RF	2,398/441	619/315	221/169	89/70	51/0	0
SENTIMETRIX-RF	2,541/443	154/0	93/0	25/0	14/0	531
SENTIMETRIX-DBSCAN	2,157/2,075	551/5,332	271/209	92/118	72/696	215

This work was supported through DoD Minerva program and AFOSR (grant FA9550-15-1-0159).