Detecting Camouflaged Content Polluters

Liang Wu[†], Xia Hu[‡], Fred Morstatter[†], Huan Liu[†]
[†]Computer Science and Engineering, Arizona State University, USA
[‡]Department of Computer Science and Engineering, Texas A&M University, USA
{wuliang, fred.morstatter, huanliu}@asu.edu, xiahu@tamu.edu

Abstract

The connectivity and openness of the Internet have cultivated a blistering expansion of online media websites. However, the culture of openness also makes the emerging platforms an effective channel for content pollution, such as fraud, phishing, and other online abuses. To complicate the problem, content polluters actively manipulate the characteristics of the Internet through establishing links with normal users and blending the malicious information with legitimate content. The manipulated links and content, being used as camouflage, make it very intricate to detect content polluters. Recent work has investigated camouflaged fraud in networks. However, due to the lack of availability of label information for camouflaged content, it is challenging to detect content polluters with traditional approaches. In this paper, we make the first attempt on detecting camouflaged content polluters. In order to evaluate the proposed approach, we conduct experiments on real-world data. The results show that our method achieves better results than existing approaches.

Introduction

Motivated by the monetary rewards, content polluters, which include fraudsters, scammers, and spammers, unfairly overpower normal users by spreading disinformation (Wu et al. 2016), which undermines the role of Internet media in sustaining a society as a collective entity. An emerging characteristic that further complicates the problem is the *camouflage*. Due to the openness of Internet media, it is easy for content polluters to copy a significant portion of content from normal users. The polluting content that is camouflaged by the legitimate messages can be very deluding due to the *cognitive inertia*: once many genuine posts from a fraudster establish trust, the fraudulent post is likely to convince many of the readers.

Recent studies have investigated the camouflage of fraudsters from the perspective of network structures (Hooi et al. 2016; Wu et al. 2017), proving that network camouflage could be efficiently detected through studying the abnormality of the density of a graph caused by the camouflage links. In this work, we focus on precisely the other side of the problem, *i.e.,* detecting content polluters in the presence of camouflage. In order to illustrate the problem, we show a toy example in Figure 1, where a normal user posts A, B, and C, and the adversarial rival copies them to camouflage the polluting post D. Our goal is to detect content polluters in the presence of camouflage.

It is particularly difficult and challenging to detect camouflaged content polluters. Due to the massive amount of content information on Internet media, there is a lack of availability of label information for camouflaged posts. Another challenge is data scarcity. Since camouflage can take up the majority of content from a content polluter, it is not easy to identify the scarce polluting evidence, and manually labeling it could be labor-intensive.

In order to tackle the challenges, we propose to utilize label information of accounts. Account labels are easier to obtain and publicly available at a relatively large scale on various platforms. Motivated by results of recent studies that camouflage tends to be *random* while malicious content is *alike* due to the similar fraudulent targets (Hooi et al. 2016), we assume that the *intersection* of content polluters' posts in the feature space is more likely to be a signal of polluting content. Hence, we aim to investigate how Camouflaged Content Polluters can be detected with Discriminant Analysis. In particular, we introduce a novel method CCPDA, which effectively detects content polluters by mining signals of camouflaged pollution. Major contributions of this work are summarized below,

- Formally define the problem of detecting camouflaged content polluters;
- Propose a novel method CCPDA to efficiently detect camouflaged content polluters; and
- Conduct extensive experiments to evaluate the effectiveness and efficiency of CCPDA.

Problem Statement

Let $\mathbf{A} = [\mathbf{V}, \mathbf{P}, \mathbf{t}]$ be a target account set with post information $\mathbf{V}$, user-post mapping $\mathbf{P}$ and identity labels $\mathbf{t}$. The data matrix $\mathbf{V} \in \mathbb{R}^{m \times n}$ is the post information of all users, where *m* is the number of posts and *n* is the number of textual features extracted from the posts. We denote the user-post association as $\mathbf{P} \in \mathbb{R}^{u \times m}$, where *u* is the number of users. $\mathbf{P}_{i,j}$ equals to 1 if the j^{th} post is posted by the i^{th} account and equals to 0 otherwise. $\mathbf{t} \in \{0, 1\}^{u \times 1}$ records

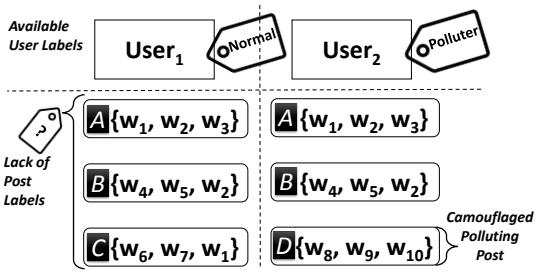

Figure 1: A toy example of camouflaged content polluters, where a normal user's posts (A, B and C) are copied to camouflage a polluting post (D).

the identity label of all users, where $\mathbf{t}_i = 1$ represents the i^{th} account is a content polluter.

We now define the problem of detecting camouflaged content polluters as follows:

Given a set of accounts $\mathbf{A}$ *with post information* $\mathbf{V}$*, user-post mapping matrix* $\mathbf{P}$*, and identity label information* $\mathbf{t}$ *for partial accounts, our goal is to learn a model with the best performance to classify whether a user is a content polluter.*

Detecting Campuflaged Content Polluters

Modeling Content Information

We represent posts with a data matrix $\mathbf{V} \in \mathbb{R}^{m \times n}$, where each row represents a post and each column represents a textual feature. However, since labels of posts are unavailable, we start with a trivial solution that labels all posts of a known polluter as polluting. Then it can be reduced to the least square problem:

$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{V}\mathbf{w} - \mathbf{y}||_2^2 + \frac{\lambda_1}{2}||\mathbf{w}||_2^2, \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^n$ is the model to learn and a regularization term $||\mathbf{w}||_2^2$ is imposed to avoid overfitting. The parameter λ_1 controls the extent of the model complexity. The vector $\mathbf{y} \in \mathbb{R}^m$ is the pseudo label that is temporally initialized. The pseudo label vector can be derived from the account labels $\mathbf{y} = \mathbf{P}^T\mathbf{t}$. However, posts of content polluters are not necessarily fraudulent, so labeling all posts of a content polluter as positive would make the classifier lose sensitivity to content pollution and result in a low *recall*.

Detecting Camouflaged Polluters with Discriminant Analysis

In order to allow for the modification of the label values, we introduce a weighting vector $\mathbf{c} \in \mathbb{R}^m$ for the label vector. Through incorporating the weight, the label of the i^{th} post becomes $c_i y_i$. So the labels could be updated through updating the weights. Our aim is to filter out camouflage, *i.e.*, increasing weights of polluting posts and decreasing weights of labels of polluters' legitimate content. To this end, we reformulate the objective function in Eq.(1) as:

$$\min_{\mathbf{w},\mathbf{c}} \frac{1}{2}\sum_{i=1}^{m}(c_i y_i - \mathbf{V}_{i,*}\mathbf{w})^2 + \frac{\lambda_1}{2}||\mathbf{w}||_2^2, \tag{2}$$

where c_i represents the weight of i^{th} post. Since the normal posts of a content polluter are initially labeled as positive, which can be viewed as mislabeled examples, they are more likely to cause a larger reconstruction error during training. Therefore, penalizing large errors leads to downweighting labels of legitimate content. In addition, since labels of legitimate users are of value 0, the weight does not influence normal users during the optimization.

Since content pollution may only comprise a small portion of all posts, the representation of $\mathbf{c}$ should be sparse. Motivated by sparse representation learning, where only few *coefficients* are assumed to reveal the key information, we incorporate an ℓ_1-$norm$ with $\mathbf{c}$ and reformulate the objective function as follows:

$$\min_{\mathbf{w},\mathbf{c}} \frac{1}{2}\sum_{i=1}^{m}(c_i y_i - \mathbf{V}_{i,*}\mathbf{w})^2 + \frac{\lambda_1}{2}||\mathbf{w}||_2^2 + \lambda_2||\mathbf{c}||_1, \tag{3}$$

where the ℓ_1-$norm$ penalizes non-sparse solutions. The parameter λ_2 controls the extent of sparsity, which can be regarded as the discriminant threshold for a post to be selected and labeled as polluting. Through introducing the sparsity regularizer, the selected entries are likely to be 1 while the unselected entries are likely to be exactly zero, which is favorable since a post is either fraudulent or legitimate in real-world applications.

Sparse representation methods are used to find dominant signals. However, some polluters would be ignored if too few polluting posts are present. In order to fully exploit the label information, we force every polluter to be selected with some posts by introducing an $\ell_{1,2}^{\mathcal{G}}$-$norm$ term. The regularization term is as follows,

$$\ell_{1,2}^{\mathcal{G}}(\mathbf{c}) = \sum_{g \in \mathcal{G}} ||\mathbf{c}_{\mathcal{G}_g}||_1^2. \tag{4}$$

The $\ell_{1,2}^{\mathcal{G}}$-$norm$, which is also called group exclusive penalty (Kong et al. 2014), is proposed to select discriminant features of different groups. Here, $\mathcal{G}$ is the set of all groups, where $\mathcal{G}_g$ denotes the indices of posts in a group $g \in \{1, 2, \ldots, m\}$. For example, let $\mathcal{G}_g = \{1, 2, 4, \cdots\}$, then $||\mathbf{c}_{\mathcal{G}_g}|| = [c_1, c_2, 0, c_4, 0, \ldots, 0]$. The $\ell_{1,2}^{\mathcal{G}}$-$norm$ first sums up absolute values of intra-group variables and then imposes an ℓ_2-$norm$ to regularize the sum. The minimization process leads to intra-group sparsity. Concretely, it enforces locally discriminant posts of a polluter to be upweighted while enforces globally discriminant content to be downweighted.

The group exclusive penalty is convex but non-smooth, which is difficult for optimization. In order to solve the problem, we rewrite the $\ell_{1,2}^{\mathcal{G}}$-$norm$ as follows (Wu, Hu, and Liu 2016),

$$\ell_{1,2}^{\mathcal{G}_g}(\mathbf{c}) = \frac{1}{2}\sum_{i=1}^{u}(\mathbf{c}^T\mathbf{P}_{i,*})^2 \tag{5}$$

$$= \frac{1}{2}\sum_{i=1}^{u}\mathbf{c}^T\mathbf{P}_{i,*}^T\mathbf{P}_{i,*}\mathbf{c} \tag{6}$$

$$= \frac{1}{2}\mathbf{c}^T\mathbf{P}^T\mathbf{P}\mathbf{c}, \tag{7}$$

where $\mathbf{P}$ denotes the user-post mapping matrix. Since $\mathbf{P}$ is a constant matrix, we introduce $\mathbf{M} = \mathbf{P}^T\mathbf{P}$ to replace the product. In $\mathbf{M} \in \mathbb{R}^{m\times m}$, $M_{i,j}$ equals to one if post i and j are generated by the same user and zero otherwise. By rewriting the regularization term $\ell_{1,2}^{\mathcal{G}_g}(\mathbf{c})$, it is convex and smooth, which can be easily incorporated into the objective function in Eq.(3) as:

$$\frac{1}{2}\sum_{i=1}^{m}(c_i y_i - \mathbf{V}_{i,*}\mathbf{w})^2 + \frac{\lambda_1}{2}||\mathbf{w}||_2^2 + \lambda_2||\mathbf{c}||_1 + \frac{\lambda_3}{2}\mathbf{c}^T\mathbf{M}\mathbf{c}, \quad (8)$$

where the parameter λ_3 controls the importance of locally discriminant content.

Optimization

The optimization problem in Eq.(8) is not jointly convex with respect to the two variables $\mathbf{w}$ and $\mathbf{c}$ together. However, by fixing one of them, the objective function is convex to the other. So we propose to find optimal solutions through alternatively updating one by fixing the other.

1)*While fixing* $\mathbf{c}$, *update* $\mathbf{w}$: the problem only depends on $\mathbf{w}$. By only considering items related to $\mathbf{w}$, we reformulate the objective as follows:

$$\epsilon_{\mathbf{w}} = \frac{1}{2}\sum_{i=1}^{m}(c_i y_i - \mathbf{V}_{i,*}\mathbf{w})^2 + \frac{\lambda_1}{2}||\mathbf{w}||_2^2, \quad (9)$$

which is reduced to an ℓ_2 regularized linear regression problem. In order to cope with the massive amount of content information, we adopt Stochastic Gradient Descent (SGD) to solve the optimization problem. SGD belongs to a class of hill-climbing optimization technique that seeks a stationary point of a function. To utilize SGD, we derive the gradient of $\mathbf{w}$ as follows:

$$\frac{\partial \epsilon_{\mathbf{w}}}{\partial \mathbf{w}} = \sum_{i=1}^{m}\mathbf{V}_{i,*}^T(\mathbf{V}_{i,*}\mathbf{w} - c_i y_i) + \lambda_1\mathbf{w}. \quad (10)$$

Instead of updating in a batch mode, SGD randomly selects data examples from the total m data instances. The update process can then be significantly accelerated with the multi-threading manner.

Therefore, the optimal predictor can be achieved through the following update rules:

$$\mathbf{w} = \mathbf{w} - \tau\frac{\partial \epsilon_{\mathbf{w}}}{\partial \mathbf{w}}, \quad (11)$$

where τ is a learning rate which we set using backtracking line search.

2)*While fixing* $\mathbf{w}$, *update* $\mathbf{c}$: the problem only depends on $\mathbf{c}$. Since the reconstruction $\mathbf{V}_{i,*}\mathbf{w}$ becomes constant, we use $\mathbf{e}$ to replace it, where $e_i = \mathbf{V}_{i,*}\mathbf{w}$. Thus, Eq.(8) can be reformulated as follows:

$$\min_{\mathbf{c}} \frac{1}{2}\sum_{i=1}^{m}(c_i y_i - e_i)^2 + \lambda_2||\mathbf{c}||_1 + \frac{\lambda_3}{2}\mathbf{c}^T\mathbf{M}\mathbf{c}. \quad (12)$$

Though all components in Eq.(12) are convex with respect to $\mathbf{c}$, the ℓ_1-$norm$ makes it non-smooth, which is difficult to optimize. Following (Liu, Ji, and Ye 2009), we try to optimize the problem in Eq.(12) through reformulating it as an equivalent smooth and convex problem,

$$\begin{aligned} \min_{\mathbf{c}\in\mathcal{Z}} \quad & \mathcal{O}(\mathbf{c}) = \frac{1}{2}||\mathbf{c}\circ\mathbf{y} - \mathbf{e}||_2^2 + \frac{\lambda_3}{2}\mathbf{c}^T\mathbf{M}\mathbf{c}, \\ \text{where} \quad & \mathcal{Z} = \{\mathbf{c} \mid ||\mathbf{c}||_1 \le z\}. \end{aligned} \quad (13)$$

$\circ$ denotes component-wise multiplication. $z \ge 0$ is the radius of the ℓ_1-*ball*. λ_2 and z have a 1:1 correspondence.

The ℓ_1-*ball* constrained convex problem in Eq.(13) can be efficiently solved. Motivated by (Ji and Ye 2009), we adopt proximal gradient descent in this work. The update rule for $\mathbf{c}$ can be formulated as follows:

$$\mathbf{c}^t = \arg\min_{\mathbf{c}\in\mathcal{Z}} P_{\gamma,\mathbf{c}^{t-1}}(\mathbf{c}), \quad (14)$$

where the superscript t denotes the number of iteration, and $P_{\gamma,\mathbf{c}^{t-1}}(\mathbf{c})$ is the convex problem's Euclidean projection onto the constraint space. The projection can be formulated as follows,

$$P_{\gamma,\mathbf{c}^{t-1}}(\mathbf{c}) = \mathcal{O}(\mathbf{c}^{t-1}) + \langle\nabla\mathcal{O}(\mathbf{c}^{t-1}), \mathbf{c}-\mathbf{c}^{t-1}\rangle + \frac{\gamma}{2}||\mathbf{c}-\mathbf{c}^{t-1}||_2^2, \quad (15)$$

where $\nabla\mathcal{O}(\cdot)$ is the derivative of $\mathcal{O}(\cdot)$. Since $\mathcal{O}(\cdot)$ is convex, $\nabla\mathcal{O}(\cdot)$ can be derived from Eq.(13) as

$$\nabla\mathcal{O}(\mathbf{c}) = \mathbf{y}\circ\mathbf{c}\circ\mathbf{y} - \mathbf{y}\circ\mathbf{e} + \lambda_3\mathbf{M}\mathbf{c}. \quad (16)$$

Given a problem in the form of Eq.(15), the analytical solution can be directly obtained (Ji and Ye 2009). The solution of $\mathbf{c}$ can be written as

$$c_j^t = max(0, u_j^{t-1}(1 - \frac{\lambda_3}{\gamma|u_j^{t-1}|})), \quad (17)$$

where $\mathbf{u}^t = \mathbf{c}^t - \frac{1}{\gamma}(\nabla\mathcal{O}(\mathbf{c}^t))$, which is introduced to replace the gradient step, and u_j^t and c_j^t are the j^{th} element of $\mathbf{u}^t$ and $\mathbf{c}^t$, correspondingly.

Table 1: Statistics of the dataset used in this study.

Posts	Reposts	Unique Users	Positive Ratio
1,150,192	576,167	94,535	7.5%

Experiments

Dataset

Existing studies obtain normal accounts through random sampling, where the cutoff between positive and negative examples may not be reflective of the original data. In order to keep in line with the real world distribution, we build up a dataset by randomly crawling all accounts under certain topics, where labels are obtained using the gold standard. In particular, we randomly sample posts from Twitter in 2013. In May 2016, we crawl each user in the dataset again and check the account status. Statistics are shown in Table 1.

Table 2: Results on Twitter dataset.

Method	Precision	Recall	F-score
SVM	29.24%	8.78%	13.53%
GBDT	69.51%	5.12%	9.66%
AdaBoost	73.41%	8.24%	14.82%
SSDM	88.01%	10.11%	18.14%
SVM_P	18.53%	62.14%	28.54%
GBDT_P	55.34%	32.22%	40.72%
AdaBoost_P	27.62%	31.85%	29.59%
SVMIL	84.36%	13.56%	23.36%
CCPDA	89.17%	30.26%	**45.96%**

Settings

We include two kinds of baselines: account-centric and post-centric methods. Account-centric methods conventionally construct an attribute vector from all posts of a user. Post-centric methods model a user by learning individual posts. The 10-fold cross-validation is employed to generate all experimental results. We list all baseline methods below,

Support Vector Machines (SVMs): are supervised learning tools for solving binary classification, which have been successfully applied to various tasks.

AdaBoost is a general boosting framework. It builds up classifiers by ensembling weak classifiers.

Gradient Boosted Decision Tree is a boosting algorithm which produces a prediction model in the form of an ensemble of multiple decision trees.

SVMIL belongs to Multiple Instance Learning (MIL) algorithms which extends SVM in a multi-instance setting. MIL shares a similar formulation with our work, assuming that each example contains multiple instances (Zhou 2004).

Social Spammer Detection in Microblogging: Hu et al. proposed a framework SSDM to detect content polluters in social media by jointly modeling network and content information (Hu et al. 2013).

Post-Centric methods are trained with individual posts. The method is then named by adding a subscript of *P* to that of corresponding account-centric models. Since post labels are not available, known content polluters' posts are all labeled as positive.

Experimental Results

The results are summarized in Table 2. Based on the experimental results, we have the following observations:

1) Post-centric methods achieve better recall while account-centric methods achieve better precision. Since an individual suspicious post causes an account to be classified as a content polluter, post-centric methods are more likely to detect more polluters, which results in the higher recall. By mixing all content together, account-centric approaches focus on the apparent content polluters, so it results in a higher precision.

2) GBDT_P achieves the second best F-score by capturing approximately $\frac{1}{3}$ content polluters with a 55% precision. Since the dataset is skewed, post-centric methods get better F-score by labeling more content polluters.

3) SVMIL performs better on precision while worse on recall. The assumption of multi-instance learning that positive bags share similar instances leads the model to focus more on obvious polluting content, and thus it loses the sensitivity to content polluters with locally discriminant polluting evidence.

4) CCPDA outperforms all the baselines with respect to F-score. In looking into the results of post-centric approaches, we find that they are oversensitive and have a lower precision. With discriminant analysis, CCPDA achieves a higher precision by focusing on only the polluting content.

Conclusion

Camouflage of content polluters presents great challenges to Internet media. In this work, we investigate how the camouflaged polluting signal can be identified with label information only for accounts. In particular, the proposed framework utilizes discriminant analysis to discover the key post that distinguishes content polluters. Experimental results on real-world data demonstrate that the proposed framework can effectively utilize available information to outperform the state-of-the-art approaches. In addition, conducting linguistic analysis on the detected content information to better understand motivations and behaviors of content polluters is also a promising direction.

Acknowledgments

The work is funded, in part, by ONR N00014-16-1-2257 and the Department of Defense under the MINERVA initiative through the ONR N000141310835.

References

Hooi, B.; Song, H. A.; Beutel, A.; Shah, N.; Shin, K.; and Faloutsos, C. 2016. Fraudar: bounding graph fraud in the face of camouflage. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 895–904.

Hu, X.; Tang, J.; Zhang, Y.; and Liu, H. 2013. Social spammer detection in microblogging. In *IJCAI*, volume 13, 2633–2639.

Ji, S., and Ye, J. 2009. An accelerated gradient method for trace norm minimization. In *ICML*, 457–464. ACM.

Kong, D.; Fujimaki, R.; Liu, J.; Nie, F.; and Ding, C. 2014. Exclusive feature learning on arbitrary structures via l12-norm. In *NIPS*, 1655–1663.

Liu, J.; Ji, S.; and Ye, J. 2009. Multi-task feature learning via efficient l 2, 1-norm minimization. In *UAI*, 339–348.

Wu, L.; Morstatter, F.; Hu, X.; and Liu, H. 2016. Chapter 5: Mining misinformation in social media. In *Big Data in Complex and Social Networks*. CRC Press. 123–152.

Wu, L.; Hu, X.; Morstatter, F.; and Liu, H. 2017. Adaptive spammer detection with sparse group modeling. In *ICWSM*.

Wu, L.; Hu, X.; and Liu, H. 2016. Relational learning with social status analysis. In *International Conference on Web Search and Data Mining*, 513–522. ACM.

Zhou, Z.-H. 2004. Multi-instance learning: A survey. *Technical Report,AI Lab, Nanjing University*.