Using Bi-level Penalized Logistic Classifier to Detect Zombie Accounts in Online Social Networks

Jing Deng
Dept. of Computer Science
UNC Greensboro
Greensboro, NC, USA
jing.deng@uncg.edu

Xiaoli Gao
Dept. of Math and Statistics
UNC Greensboro
Greensboro, NC, USA
x_gao2@uncg.edu

Chunyue Wang
College of Comm. Engr.
Jilin University
Changchun, Jilin, P. R. China
chunyue@jlu.edu.cn

ABSTRACT

The huge popularity of online social networks and the potential financial gain have led to the creation and proliferation of zombie accounts, i.e., fake user accounts. For considerable amount of payment, zombie accounts can be directed by their managers to provide pre-arranged biased reactions to different social events or the quality of a commercial product. It is thus critical to detect and screen these accounts. Prior arts are either inaccurate or relying heavily on complex posting/tweeting behaviors in the classification process of normal/zombie accounts. In this work, we propose to use a bi-level penalized logistic classifier, an efficient high-dimensional data analysis technique, to detect zombie accounts based on their publicly available profile information and the statistics of their followers' registration locations. Our approach, termed (B)i-level (P)enalized (LO)gistic (C)lassifier (BPLOC), is data adaptive and can be extended to mount more accurate detections. Our experimental results are based on a small number of SINA WeiBo accounts and have demonstrated that BPLOC can classify zombie accounts accurately.

CCS Concepts

•Security and privacy → Social network security and privacy;

Keywords

Online social networks; zombie detection; classfier

1. INTRODUCTION

With the high adoption rate of online social network Apps on smartphones as well as other computing platforms, there is a lot to gain if the behaviors of a larger number of accounts can be controlled. For example, these controlled accounts can be directed to praise a local event in order to snowball more positive posts/tweets toward the event. In turn, such praises can be maneuvered or "bought", masquerading the reactions from real people or crowd.

However, it is difficult to control the behavior of real people and their accounts. Instead, zombie accounts or accounts without individual personality behind can be created, maintained, and managed. Such accounts can be easily directed to perform coordinated actions such as praising a new product, trashing a competing product, or even "bought" for other larger-scale online opinion maneuvers. These accounts are usually controlled by a machine or sometimes managed by the people creating such accounts.

Due to high diversity in user behaviors, detection of such zombie accounts or distinguishing them from regular user accounts is difficult, though not impossible. There are sometimes common features that are expected in the behavior of such accounts, for instance, similarity in postings, coordinated posting times, meaningless postings that sometimes can be easier to detect by human but not algorithms or machines, and frequent reposting (or forwarding) of other hot-topic posts. Zombie accounts' behavior can be vastly different as many of these are occasionally managed by a few people behind. Their adaptive maneuver of zombie accounts posting behavior makes it even harder to detect them successfully. After all, though unlikely, nothing stops them to suddenly behave as real users for some time. Similarly, there are abandoned real user accounts with users having moved away and hacked accounts whose activities had been normal until a certain time, making detection more challenging.

In this work, we focus on the detection of zombie accounts in twitter-like online social networks. We used SINA WeiBo as the platform for our experiments. SINA WeiBo, with the meaning of microblogging, is one of the most popular online social networks in China. Launched around 2009, WeiBo has quickly become one of the most significant and active information dissemination centers. Many of the posts focus on entertainment, breaking news, opinions, as well as commercial products. The gigantic number of registered accounts has been questioned, as it has been once reported that more than 50% of the accounts in WeiBo had empty timelines [20][13][12][9].

Our method, termed Bi-level Penalized LOgistic Classifier (BPLOC), is to look into a list of profile information with natural grouping features including number of followers, number of friends (followings), number of bi-followers, as well as registration location information (province/state and city). Although users can claim whatever registration location during the registration process or even change it at a later time, the selections at least represent users' preferred

ICNCC '16 December 17–21, 2016, Kyoto, Japan

DOI:

location for online interaction. BPLOC is able to select important variables to detect zombie accounts accurately. As a by-product, BPLOC also explains how important features affect the odds of an account being a zombie account.

The remainder of this paper is structured as follows. Section 2 presents related work; Section 3 discusses mathematical background and our technique; Section 4 shows the performance of our technique; finally, Section 5 concludes this paper.

2. RELATED WORK

There have been many research on online social networks. Some of these investigated on those accounts with no real users behind them. For instance, Thomas et al. [17] looked into suspended Twitter accounts and analyzed their lifetime events and behaviors. In [3], Chu et al. focused on contents instead of individual account behaviors. Ghosh et al. investigated link farming issue and proposed to collusionRank to deter users to follow potential spammers on Twitter [5].

There have been works focusing on zombie account detection in Twitter-like online social networks including SINA WeiBo. Shen et al. [16] used a binary classifier based on several major features, which include number of followers, number of friends (followings), posting behavior during daytime and nighttime, as well as changes of posting behavior over time. Users' posting behaviors were inspected by Guo et al. in [6]. Some used social interactions among different accounts to identify fake user accounts [19][8]. In [18], word distribution of suspected accounts' comments toward highly influential verified accounts' tweets are tested before account classification is made. Other approaches include finding synchronous behavior among zombie users [10], large-scale user set and posting record [1], two-stage cascading detection model [11], subgraph analysis [2], and registration location [13][12][4].

3. BI-LEVEL PENALIZED LOGISTIC CLASSIFIER (BPLOC)

3.1 Profile variables

Each WeiBo user has a list of different profile variables, most of which can be retrieved through WeiBo API. Among these variables, city and province need some more explanation: from our data, we found that there were 36 provinces[1] and 550 different registered cities in SINA WeiBo. We modeled these as 0/1-valued variable for each user account. Therefore, if an account claimed registration location as, e.g., Guangdong, Foshan, we mark 1 on province Guangdong and 1 on Guangdong, Foshan and all other province and city variables with values of 0. There are altogether 596 variables (36 provinces, 550 cities, and 10 other variables.[2]

The SameC.r and SameP.r variables are unique in our data collection. Essentially, these represent the ratio of one's followers who registered from the same city (SameC.r) and province (SameP.r), respectively.

$$\text{SameC.r} = \frac{\text{\# of followers residing in the same city}}{\text{Total \# of followers retrieved}}$$

and

$$\text{SameP.r} = \frac{\text{\# of followers residing in the same province(state)}}{\text{Total \# of followers retrieved}}$$

And these variables take values between 0 and 1. Note that when two accounts have the same registration city, it also means that they have the same registration province/state. Hence,

$$\text{SameC.r} \leq \text{SameP.r} \tag{1}$$

As reported in [4], such ratios are important in the detection of zombie accounts because real users interact with friends or colleagues who are geographically close to them. Instead, zombie accounts are expected to have evenly distribute registration locations in order to avoid easy detection from high concentration at some special locations.

3.2 Model setting

Logistic regression is a predictive modeling technique in classification by predicting odds of an event for all subjects from certain features. The logistic classifier has become quite popular in classification [7].

Let y_i be the zombie status. In particular, let $y_i = 1$ when account i is a zombie, and 0 otherwise. Denote x_{ij}, $j = 1, \cdots, p$, its p features. A logistic regression is to predict the natural logarithm of odds being a zombie by

$$\log\left(\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right) = \beta_0 + \eta_\beta(\mathbf{x}_i) \ , \tag{2}$$

where $\eta_\beta(\mathbf{x}_i) = \sum_{j=1}^{p} x_{ij}\beta_j$, β_j, $j = 1, \cdots, p$, are the estimates, and β_0 is the intercept.

In high-dimensional data setting, when the numbers of features is large compared to the sample size, penalized logistic regression is often used for simultaneous feature selection and classification [15][14]. Since multiple city and province information are included in our zombie data, those cities and provinces may function group-wise in predicting the zombie status. We adopt a bi-level group-wise variable selection in logistic regression.

Suppose these p features x_{ij}, $j = 1, \cdots, p$, can be divided into J groups, with K_j elements in group j: β_{jk} for $k = 1, \cdots, K_j$. It is reasonable to assume that only some groups are important features for zombie status classification, i.e., the contribution of all features to zombie classification is sparse at the group level. In addition, for those important groups, only a few members are truly relevant to zombie classification. Thus, the feature's importance is sparse within group level.

A BPLOC is to obtain both feature selection and classification using the MCP (Minimax Concave Penalty) criteria [21] at both between-group and within-group levels. In particular, we estimate β_j, $j = 1, \cdots, p$, by maximizing the following penalized maximum likelihood,

$$\text{PL}(\beta, \lambda) = l(\beta) - \sum_{j=1}^{J} f_{\lambda,b}\left(\sum_{k=1}^{K_j} f_{\lambda,a}(|\beta_{jk}|)\right) \ , \tag{3}$$

[1]Probably with the purpose of categorizing registered users in a more evenly fashion, SINA WeiBo counts abroad as a province and different foreign countries as cities.

[2]Note that our limited training data provided about 135 variables and we leave a more thorough study of all different locations and larger training size to our future work.

where $b = 3$

$$l(\beta) = \sum_{i=1}^{n} y_i \eta_\beta(\mathbf{x}_i) - \log[1 + \exp\{\eta_\beta(\mathbf{x}_i)\}] \quad (4)$$

and

$$f_{\lambda,a}(\theta) = \begin{cases} \lambda\theta - \theta^2/(2a) & \text{if } \theta \le a\lambda \\ 0.5a\lambda^2 & \text{if } \theta > a\lambda \end{cases}, \quad (5)$$

where $a = 3$ and λ is a tuning parameter. λ is important in controlling the bi-level sparsity. The larger λ is, the fewer features we will select to use for generating the final logistic classifier. We first use cross validation to select an optimal λ, and then use the training data to generate a post-selection logistic classifier.

4. PERFORMANCE EVALUATION

Performance evaluation of our scheme is based on the data we collected from SINA WeiBo in late 2014. Our script crawled from one WeiBo account and walked through all the (publicly listed) followers of this account. Then all these followers' (publicly listed) followers are crawled. Our script stopped after having collected 10,000 inspected accounts. For each of these inspected accounts, our script retrieved all of the public profile information, including number of followers, registration date, registration province/state, registration city, number of followings, number of friends(followers), favourites_count, statuses, number of bi-followers (those accounts who follow this account and are followed by this account), as well as other information. Then our script walked through all of the shown followers and collected their profile information as well. Altogether, it collected public profile information of about 800,000 accounts.

Training data were collected through the following process: we randomly selected about 130 from the 10,000 accounts and checked objectively whether these were real user accounts or zombie accounts based on all information we could read including posting frequency, posting timing, actual posts/tweets, etc.

Performance metrics include false positive (real accounts detected as zombie accounts), false negative (zombie accounts detected as real accounts), and F-score, which can be defined as

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \quad (6)$$

where precision is the rate measured as true positive divided by all positives and recall is successful detection rate (rate of all detected zombies divided by all zombies).

Figure 1 shows the solution path of group MCP. Each line except the gray vertical one represents how the coefficient of a variable changes with λ. The gray vertical line identifies the location of optimal λ. The sequence of the curves crossing the horizontal line from left to right is the sequence that the variables being added to the set of selected variables.

Based on Figure 1, SameC.r and favourites_count enter our detection model quite early with small λ. It means that these variables are important in our classifier model. Favourites_count is the number of saved tweets of other accounts. A larger number means that the account is less likely to be a zombie. With respect to SameC.r, regular active accounts usually would interact with some accounts from the same city. Therefore, a higher SameC.r means

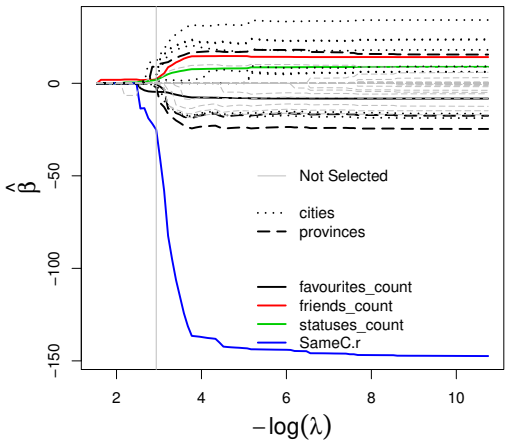

Figure 1: Variable Selection as λ changes.

Table 1: Coefficients Estimates and Odds Ratio

variable name	Estimate	Odds
favourites_count	-5.795	3.041×10^{-3}
friends_count	4.666	106.294
statuses_count	4.153	63.595
SameC.r	-59.664	0.551
city.f110006	20.522	8.175×10^{8}
city.f220005	26.419	2.977×10^{11}
city.f330003	26.472	3.139×10^{11}
city.f330004	20.278	6.406×10^{8}
city.f410008	-20.738	9.858×10^{-10}
city.f430004	-23.692	5.139×10^{-11}
city.f440003	11.426	9.169×10^{4}
city.f500083	21.320	1.816×10^{9}
city.f4001000	30.882	6.465×10^{5}
province.f14	13.379	2.581×10^{13}
province.f31	-23.729	4.952×10^{-11}
province.f82	-21.058	7.153×10^{-10}

that the account is less likely to be a zombie. Our classifier suggests that a higher friends_count means a higher chance of the account being a zombie. This is because of zombie accounts' aggressive following strategy (hoping for a follow-back). Statuses_count is basically the number of tweets: zombie accounts tweet aggressively.

It seems that some cities and provinces/states are more likely to be chosen as registered locations for zombie accounts. This is either because of their higher popularity in WeiBo users, training data randomness, or their subtle attraction to zombie accounts for popularity, requiring further investigation that is out of the scope of this work.

The estimated coefficients and their corresponding odds ratios are given in Table 1. [3] [4] The last column of Table 1 is the corresponding odds ratio computed from the coefficient

[3]Note that the odds ratio for SameC.r is under 1% change.
[4]Note that all count numbers have a unit of 1000, except favourites_count with a unit of 10.

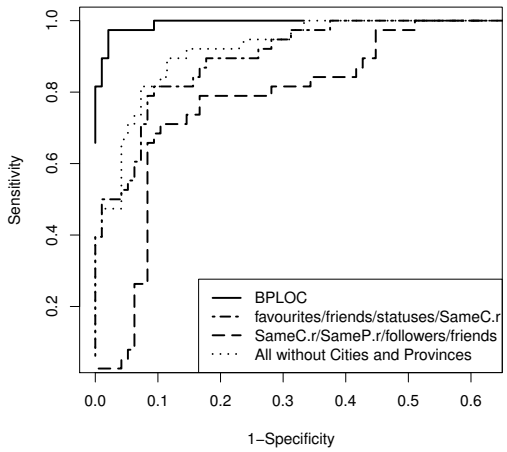

Figure 2: ROC Curve of different classifiers.

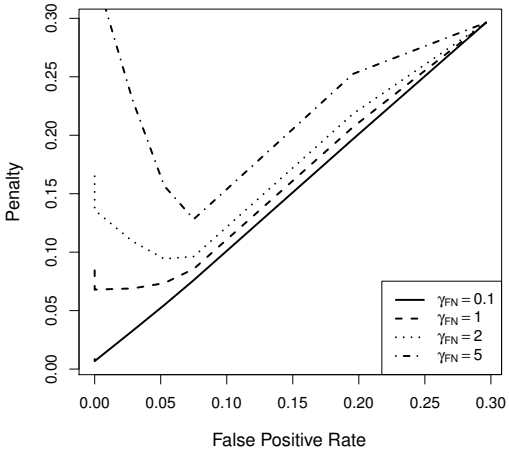

Figure 3: Penalty minimization

estimation. For example, if an account's favourites_count increases by 10, the odds of being a zombie reduces by 99.7%; if the friends_count increases by 1000, the odds of being a zombie increases 106 times; if the SameC.r increases by 1%, the odds of being a zombie reduces by about 45%. Since we used city.f110001 (Beijing, Dongcheng District) and province.f12 (Tianjin) as the baseline regions, if an account is registered in city.f110006 (Beijing, Fengtai District), the odds of being a zombie is 8.175×10^8 times that of an account registered in city.f110001. Similarly, the odds for an account in province.f82 (Macau) being a zombie is 7.153×10^{-10} times of one in province.f12.

Figure 2 presents the ROC (receiver operating characteristic) curve of our classifier model. Based on Figure 2, we can see that, it is rather inaccurate to use only SameC.r, SameP.r, number of followers, and number of friends as the classifier variables (dashed line). Dotted-dashed line is slightly better, using four other variables: favourites_count, number of friends, number of tweets, and SameC.r. The dotted line presents the results for the classifier using all variables but not registration city and province information. The best ROC performance belongs to BPLOC, operating on top-left corner, which is the best region to operate upon with high true positive rate and low false positive rate.

Suppose different penalties are associated with false positives γ_{FP} and false negatives γ_{FN} (examples include a certain monetary penalty of making false positive decisions and a different penalty of making false negative decisions), the overall penalty is then

$$\Gamma = \gamma_{FP} \cdot r_{FP} + \gamma_{FN} \cdot r_{FN} \tag{7}$$

The results of minimized penalty Γ is shown in Figure 3. Without loss of generality, we assume $\gamma_{FP} = 1$ and present results for different false negative penalty, $\gamma_{FN} = 0.1, 1, 2, 5$. In Figure 3, most of the curves are convex and show false positive rates where the classifier should work on in order to achieve the lowest penalty. As γ_{FN} increases, the best false positive rate associated with γ_{FN} increases.

In Figure 4, we show the F1-score of the BPLOC scheme with different cutoff values. The 0.8 F1-score achieved with a cutoff value of 0.4 underlines the strong performance of BPLOC.

5. CONCLUSIONS

Online social networks have seen the rise of many fake users, zombie accounts, due to potential monetary gains. Because of the vague differences among zombie accounts, real user accounts, abandoned accounts, idle accounts, and hacked accounts, it is excessively difficult to identify them and many different approaches have been designed and investigated. In this work, we have proposed a scheme called Bi-level Penalized Logistic Classifier (BPLOC). The BPLOC scheme incorporates group information in the process of high-dimensional variable selection and has been demonstrated to be powerful as well as highly accurate in classification. The BPLOC scheme does not require complex posting behavior or post information before account classification is performed.

Using BPLOC, we have confirmed the importance of same city ratio of an account's followers as has been highlighted in

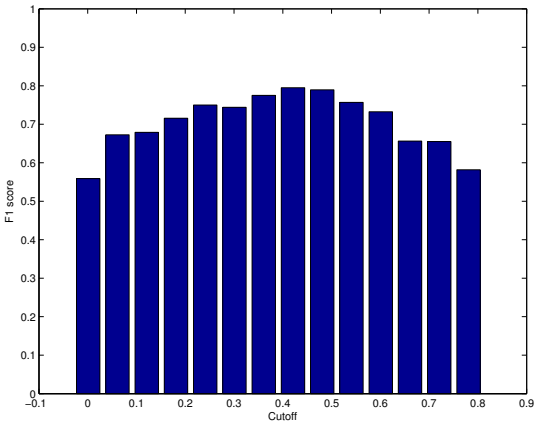

Figure 4: F1 score for different cutoff values

[20]. In addition, we have addressed the importance of several other features including favourites count, friends count and statuses count. We believe that, compared to other approaches requiring user posting records or past tweets/posts, the requirement of mostly publicly available follower profile information by BPLOC is reasonable and necessary so that more accurate detections can be made.

In our future work, we will look into registration date and time in order to improve our classifier's accuracy. A significantly larger-sized training data would further help to increase BPLOC's classification accuracy as well.

Acknowledgment

Deng's research is in part supported by NSF grant CCF-1320428. Gao's research is in part supported by Simons Foundation (359337).

6. REFERENCES

[1] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, pages 1–9, 2010.

[2] A. Beutel, L. Akoglu, and C. Faloutsos. Fraud detection through graph-based user behavior modeling. In *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pages 1696–1697, New York, NY, USA, 2015. ACM.

[3] Z. Chu, I. Widjaja, and H. Wang. Detecting social spam campaigns on twitter. In F. Bao, P. Samarati, and J. Zhou, editors, *Applied Cryptography and Network Security*, volume 7341 of *Lecture Notes in Computer Science*, pages 455–472. Springer Berlin Heidelberg, 2012.

[4] J. Deng, L. Fu, and Y. Yang. ZLOC: Detection of zombie users in online social networks using location information. In *Proc. of the third IARIA International Conference on Building and Exploring Web Based Environments (WEB 2015)*, pages 24–28, Rome, Italy, May 24-29 2015.

[5] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi. Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 61–70, New York, NY, USA, 2012. ACM.

[6] Z. Guo, Z. Li, H. Tu, and L. Li. Characterizing user behavior in weibo. In *Mobile, Ubiquitous, and Intelligent Computing (MUSIC), 2012 Third FTRA International Conference on*, pages 60–65, June 2012.

[7] T. Hastie, R. Tibshirani, and J. H. Friedman. *Elements of Statistical Learning: Data Mining, Inference and Prediction (Second Edition)*. Springer-Verlag, New York, 2009.

[8] X. Hu, J. Tang, Y. Zhang, and H. Liu. Social spammer detection in microblogging. In *23rd International Joint Conference on Artificial Intelligence (IJCAI '13)*, August 3-9 2013.

[9] H. Jiang, Y. Wang, and M. Zhu. Discrimination of zombie fans on weibo based on features extraction and business-driven analysis. In *Proceedings of the 17th International Conference on Electronic Commerce 2015*, ICEC '15, pages 13:1–13:5, New York, NY, USA, 2015. ACM.

[10] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang. Detecting suspicious following behavior in multimillion-node social networks. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 305–306, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.

[11] H. Liu, Y. Zhang, H. Lin, J. Wu, Z. Wu, and X. Zhang. How many zombies around you? In *Proc. on Data Mining (ICDM), IEEE 13th International Conference on*, pages 1133–1138, Dec 2013.

[12] J. Liu, W. Yu, and S. Li. A framework to extract keywords from sina weibo data for tracking user trail. *Journal of Information and Computational Science*, 12(1):51–58, January 2015.

[13] J. Lu, X. Yu, and W. Wan. Analysis of topology and properties on localized microblog network. *Journal of Information and Computational Science*, 11(10):3503–3512, October 2014.

[14] L. Meier, S. van de Geer, , and P. Buhlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 70(1):53Ű71, 2008.

[15] M. Y. Park and T. Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 69:659Ű677, 2007.

[16] Y. Shen, J. Yu, K. Dong, and K. Nan. Automatic fake followers detection in chinese micro-blogging system. In V. Tseng, T. Ho, Z.-H. Zhou, A. Chen, and H.-Y. Kao, editors, *Advances in Knowledge Discovery and Data Mining*, volume 8444 of *Lecture Notes in Computer Science*, pages 596–607. Springer International Publishing, 2014.

[17] K. Thomas, C. Grier, D. Song, and V. Paxson. Suspended accounts in retrospect: An analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '11, pages 243–258, New York, NY, USA, 2011. ACM.

[18] H. Wang, K. Lei, and K. Xu. Profiling the followers of the most influential and verified users on sina weibo. In *Communications (ICC), 2015 IEEE International Conference on*, pages 1158–1163, June 2015.

[19] X. Wu, Z. Feng, W. Fan, J. Gao, and Y. Yu. Detecting marionette microblog users for improved information credibility. In H. Blockeel, K. Kersting, S. Nijssen, and F. Zelezny, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *Lecture Notes in Computer Science*, pages 483–498. Springer Berlin Heidelberg, 2013.

[20] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, and Y. Dai. Uncovering social network sybils in the wild. *ACM Transactions on Knowledge Discovery from Data*, 8(1):2:1–2:29, Feb. 2014.

[21] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942, 2010.