Detection of Misbehaviors in Clone Identities on Online Social Networks

Rishabh Kaushal[1,2], Chetna Sharma[2], and Ponnurangam Kumaraguru[1]

[1] Precog Research Group
Indraprastha Institute of Information Technology, Delhi, India
{rishabhk,pk}@iiitd.ac.in
[2] Indira Gandhi Delhi Technical University for Women, Delhi, India
chetna712@gmail.com

Abstract. The account registration steps in Online Social Networks (OSNs) are simple to facilitate users to join the OSN sites. Alongside, Personally Identifiable Information (PII) of users is readily available online. Therefore, it becomes trivial for a malicious user (*attacker*) to create a spoofed identity of a real user (*victim*), which we refer to as *clone identity*. While a victim can be an ordinary or a famous person, we focus our attention on clone identities of famous persons (*celebrity clones*). These clone identities ride on the credibility and popularity of celebrities to gain engagement and impact. In this work, we analyze celebrity clone identities and extract an exhaustive set of 40 features based on posting behavior, friend network and profile attributes. Accordingly, we characterize their behavior as *benign* and *malicious*. On detailed inspection, we find benign behaviors are either to promote the celebrity which they have cloned or seek attention, thereby helping in the popularity of celebrity. However, on the contrary, we also find malicious behaviors (*misbehaviors*) wherein clone celebrities indulge in spreading indecent content, issuing advisories and opinions on contentious topics. We evaluate our approach on a real social network (Twitter) by constructing a machine learning based model to automatically classify behaviors of clone identities, and achieve accuracies of 86%, 95%, 74%, 92% & 63% for five clone behaviors corresponding to promotion, indecency, attention-seeking, advisory and opinionated.

Keywords: Online Social Networks · User Clone Identities · Behavioral Detection.

1 Introduction

Online Social Networks (OSNs) offer people in the real world to create accounts to avail plethora of social services being offered in the virtual world. While in the real world, it is readily feasible to verify the identity of an individual, it is quite tricky in OSNs [15]. The process of account creation is offered in quick and easy steps to encourage the adoption of OSNs platforms. This helps users create their accounts (also referred to as identities) with much ease. While it

(a) Victim (b) Clone (c) Fan (d) Other

Fig. 1: Illustration of Victim, Clone, Fan and Other Identities in Twitter.

helps genuine users create identities easily, on the flip side, it also enables a malicious user to create identity *similar* to a genuine user (victim), which we refer to as *clone identity*[3] [2]. The public availability of Personally Identifiable Information (PII) of users, like, profile picture, bio details and name, makes the task of a malicious user even more trivial [18]. We note that clone identities are different from *fake* identities (or *sybils*) in which an attacker creates a random profile without impersonating any individual.

In this work, we focus our attention on the clone identities of celebrities. The motivations for a malicious user to create clone identities are many-fold as exhibited by their behaviors. For instance, Fig 1 depicts victim (well known Indian film celebrity Amitabh Bachchan on Twitter, Fig 1a) along with his clone identity (Fig 1b), which has been in existence for a long time (since 2009 in this case). Fan identity (in case of celebrity) also exists as shown in Fig 1c along with an identity (Fig 1d), which has the same name but is neither clone nor fan. Clone identities indulge in several behaviors as depicted in Fig 2 such as promotion (Fig 2a), indecency (Fig 2b), attention-seeking (Fig 2c), advisory (Fig 2d) and opinionating (Fig 2e). In the case of *celebrity cloning* [6], the apparent motivation is to ride on the popularity and reputation of known celebrities to influence users on OSN platforms. While behaviors associated with promotion and attention-seeking are *benign*, on the other hand, the behavior of spreading indecency is undoubtedly *malicious*. Also, the behaviors that involve sending advisories and opinions, particularly on contentious issues, that misrepresent celebrities, would be considered as *malicious* behaviors. Besides celebrities, clone identities are being created for ordinary individuals as well, in order to create *real-looking* profiles. These profiles are subsequently used to launch social engineering attacks like fake-following [3, 8], fake-likes [17], spear-phishing [16]. In this work, we do not consider clones of ordinary people since their reach and impact is mostly limited to the victim alone.

[3] It is also referred as impersonation attack or identity clone attack.

(a) Promotion (b) Indecency (c) Attention Seeking

(d) Advisory (e) Opinionated

Fig. 2: Behavioral Characteristics Exhibited by Clone Identities.

Our proposed solution of behavior characterization of clone identities consists of the following steps. In the first step, we find suspected clone identities of the victim. These suspected identities are marked as *clone* identities, *fan*[4] identities (in case of celebrities), and *none*. In the second step, behavioral characterization of each of the clone identity is performed into predefined categories based on their behavior, as shown in Fig 2. Five categories are considered namely promotion, indecency, attention-seeking, advisory, and opinionated. Our *behavioral characterization model*, pre-trained on 692 clones gives accuracies of 86%, 95%, 74%, 92% & 63%, respectively.

2 Related Work

Clone identities are a particular case of fake identities in which the victim's PII are leveraged by an attacker to create real-looking identities. Detection of fake identities, referred to as Sybil attacks are well studied. SybilGuard from Yu et al. [21] examines the impact of multiple fake identities (Sybil nodes) on honest nodes. Viswanath et al. [19] summarize the design of Sybil defense space from the perspective of detecting Sybils and tolerating (quantifying) their impact. Cao et al. [7] introduce a notion of ranking nodes (*SybilRank*) regarding their likelihood of being fake. While these works leverage network-based information in their solution approaches, Wang et al. [20] explore the possibility of a crowd-sourced solution for the detection of Sybils. Gupta et al. [11] leverage the machine learning approach for the detection of fake accounts on Facebook.

[4] Fan identities are created by supporters of celebrities with benign intentions of popularizing the celebrity. They may also be created by celebrities themselves, however, we don't delve into these issues, since our key focus is on behavior of clone identities.

In the context of clone detection, proposed solutions have exploited the fact that the attacker creates clone identities with attributes similar to that of the victim. Bilge et al. [5] demonstrate an identity theft attack on existing users of a given OSN and improve the trustworthiness of these identities by sending a friend request to friends of cloned victims. In another attack, they create cloned identities of victims across other OSNs where victims did not have their presence. Jin et al. [13] exploit attribute similarity and common friends as critical indicators to find clone identities. Kharaji et al. [14] also explore the similarity of attributes and strength of relationships as essential features to detect clone identities. However, both [13] and [14] could not validate their proposed approach on real OSN platform due to unavailability of verified and their clone identities. He et al. [12] propose a scheme to protect users from identity theft attacks. Gogo et al. [10] propose a technique for the collection of impersonation attacks. Their findings suggest that these attacks are targeting even ordinary individuals to create pseudo-real fake identities to evade detection.

3 Data Collection and Ground Truth

Among the various OSN platforms, we choose Twitter to evaluate our approach for many reasons. *First*, it is a popular short message service; users read and forward the tweets instantaneously. *Second*, it provides simple steps for account creation and has among the best support for developers, so creating a clone [9] is trivial. *Third*, Twitter follows a verification process for celebrities and grant a blue colored verify badge[5] indicating verified account. For selecting celebrities,

Table 1: Distribution of Suspected Clone Identities into Three Categories namely Clones, Fans and None

Victim Account	**Clones**	**Fans**	**None**	**Total**
Narendra Modi	84	38	41	163
Shah Rukh Khan	56	11	41	108
Amitabh Bachchan	86	8	78	172
Salman Khan	23	7	42	72
Akshay Kumar	17	6	176	199
Sachin Tendulkar	107	10	70	187
Virat Kohli	79	30	20	129
Deepika Padukone	129	15	74	218
Hrithik Roshan	94	9	86	189
Aamir Khan	20	0	157	177
Total	695	134	785	1,614

[5] Verified Accounts on Twitter: https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

we use TwitterCounter[6], a web-based service to get 10,977 top influential (most followed) Twitter users spread across 227 countries. Due to computational constraints, we select the ten most influential users[7] from India. For each of them, we perform user search on Twitter using Search API[8] using various combinations of the name of user (first name only, first letter of first name + last name, both first name + last name and first name + first letter of last name). As a result, we obtain 1,614 suspected clone identities. We manually inspected each of these identities to find out whether they are indeed cloned identities or fan accounts (created to publicize or support their celebrities) or none of these. Out of 1,614 suspected clone identities, 695 were found to be clones, 134 fan identities, and the remaining 785 were neither clones nor fans, which forms ground truth for clone detection. Table 1 explains the breakup of these suspected clone identities. Further, we prepare ground truth for the behavior characterization

Table 2: Distribution of Five Behavioral Categories (C1:Promotion, C2:Indecency, C3:Advisory, C4:Opinionating, C5:Attention) among Clones and Fans

Victim Account	**C1**	**C2**	**C3**	**C4**	**C5**
Narendra Modi	8	9	7	61	27
Shah Rukh Khan	7	1	11	20	16
Amitabh Bachchan	14	3	12	28	29
Salman Khan	7	1	2	9	8
Akshay Kumar	5	0	1	12	5
Sachin Tendulkar	26	5	4	47	26
Virat Kohli	18	4	5	33	33
Deepika Padukone	27	12	10	52	42
Hrithik Roshan	19	7	9	30	28
Aamir Khan	6	0	1	6	6
Total	137	42	62	298	220

of clones and fans. Out of 829 of these identities (695 clones and 134 fans), we found that 22 of them got suspended, and 115 of them did not post even a single tweet. So, ignoring these, we focused our attention on the remaining 692 identities by manually inspecting all the tweets posted by them and engagement received. Based on the kind of content being posted, we narrowed down their behavior into *five behavioral categories* namely promotion, indecent, advisory, opinions, and attention seeking. The distribution of identities belonging to these

[6] https://twittercounter.com/pages/100/

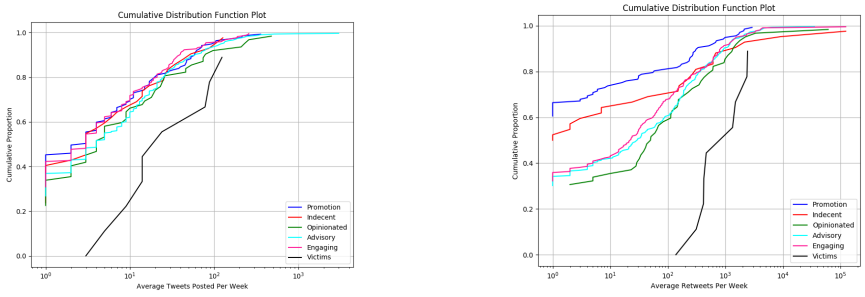
[7] Narendra Modi, Shah Rukh Khan, Amitabh Bachchan, Salman Khan, Akshay Kumar, Sachin Tendulkar, Virat Kohli, Deepika Padukone, Hrithik Roshan and Aamir Khan

[8] Twitter Search API: https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html

categories are 137, 42, 62, 298, and 220, respectively as mentioned in Table 2. We observed that all these numbers add up to 759 which means that some of these identities exhibited more than one behavior.

4 Proposed Approach

Once we have detected clones, as explained in data collection, the next step is to characterize their behavior. There are *five* behavioral categories that we focus upon namely promotion, indecent, opinionated, advisory, and attention-seeking. During our behavioral characterization study of clones, as depicted in Fig 3, we found that clones exhibit lessor activity weekly in terms of tweets posted (Fig 3a) and tweets retweeted (Fig 3b) as compared to victims who are influential users on Twitter.



(a) Average Tweets Posted Per Week (b) Avg. Retweets Received Per Week

Fig. 3: Behavioral Characteristics Exhibited by Clone Identities.

Table 3 describes the details of 40 features employed for behavioral characterization. We compute each of the features marked with '*' weekly, and we consider minimum, maximum, average, and standard deviation for each of them as features. We divide features into three categories namely content, network and profile, depending upon the type of attribute used as the source for feature computation.

– **Content Based Features**: The kind of content posted by clones provides a good indication of the type of behavior exhibited. The presence of URLs could lead users to inappropriate sites or promotional content. For instance, promotional keywords [4] would indicate promotion (or advertisement) class. Currency symbols could attract users towards some promotion. The presence of question marks and engaging words (like *who*, *what*, *when*, and*where*) could be used to invite attention or engagement. Swear words [1] would indicate the presence of indecency. Special characters like quotes and advisory

Table 3: Features for Behavioral Characterization

Features Type	**List of Features**
Content based Features (21)	URLs, Promotional Keywords, Mentions, Currency Symbols, Question Marks, Engaging Words, Swear Words, Quotes, Advisory Keywords, Days Since Last Tweet, Time* between Two Tweets, Tweet* Length, Exclamation, Colon-Semicolon.
Network based Features (14)	Tweets* per week, Retweet* Count & Favorite* Count, Followers, Following.
Profile based Features (5)	Bio Analysis - URLs, Length, Victim Tag, Fan or Clone, Mention, Handle Mention.

keywords (like *should*, *said*, and *quote*) could indicate self-help or advisory. Besides these, we use generic features like hashtags, tweet length, time between two tweets, days since the last tweet, presence of exclamation symbol and colon-semicolon.

- **Network Based Features**: Behavior of clone identities with their ego network can be studied by measuring the engagement. Therefore features like retweet count, favorite count, tweets per week, number of followers, and number of following are computed here in network-based features.
- **Profile Based Features**: Twitter has very few profile attributes among which *user bio* is worth investigating. We compute the number of occurrences of URLs, victim name (or tag) along with the length of bio in user bio field as features. Also, to capture the nature of profile as described by the user, we look into the occurrence of some common words. A clone may use words like *real account* or *official account*, whereas a fan page bio may have *unofficial page*, *parody account*, or *fan association* mentioned.

5 Evaluation and Results

We explain our evaluation methodology and corresponding results in this section. Recall from Table 2 that 692 clones (and fan) identities were analyzed to categorize them into one (or more) of the behavioral types. In particular, 137 were found to be involved in the promotion, 42 in spreading indecency, 64 in advisory, 298 in opinionating, and 220 in attention-seeking. We use this as ground truth and answer the following research questions (RQs).

- RQ1: Which is the best classifier for behavior characterization of clones?
- RQ2: Does detection accuracy improve with more training?

Identifying Best Classifier To identify the best classifier, we compute 40 features on the 692 identities and ran over 12 off-the-shelf classifiers namely Random Forest, Decision Tree, Logistic Regression, KNeighbors, ExtraTreesClassifier, Logistic Regression, Ridge Classifier, ExtraTree Classifier, Neural Network -

MLPClassifier, LinearSVC and Naive Bayes Classifier (Bernoulli and Gaussian). In our experimental set-up, we consider the multi-class (five classes) problem as five different binary classification problems in which the goal is to detect the presence or absence of a specific behavior in a given clone identity. It turns out that there is no single classifier, which performs best for all behavior types. Random forest works the best (94%) for detecting indecency, Naive-Bayes detects promotion with 86% accuracy, Logistic Regression gives 74% accuracy for attention-seeking behavior, RidgeClassifier gives 92% accuracy for advisory behavior whereas ExtraTreesClassifier gives 63% accuracy for opinionated content spreading.

Table 4: Accuracy scores with different training-testing split

Train-Test	**Promotion**	**Indecency**	**Attention Seeking**	**Advisory**	**Opinionated**
80-20	**0.86**	**0.94**	**0.92**	**0.63**	**0.74**
70-30	0.73	0.94	0.90	0.56	0.68
60-40	0.82	0.91	0.90	0.54	0.61
50-50	0.80	0.92	0.90	0.54	0.65

Training-Testing Split In this evaluation, we study the effect of train-test split on classifier performance. As evident from Table 4, the classification accuracy is improved in all behavioral types as we increase the train-test ratio from 50-50 to 80-20, which suggests that as training size would size, the accuracies will improve. Also, we observe that the accuracy of the advisory class is low due to less number of clones spreading advisory behavior (Table 2). On the contrary, the accuracy of the indecent class is high, even though the number of indecent instances is less. We attribute it to the fact that swear words in indecency are limited and highly discriminative.

6 Limitations and Future Work

There are a few limitations to this work. We carefully select Twitter as the social network platform because it provides a mechanism of *verified accounts* in which a blue tick appears in user profile. This helped us in correctly identifying the real account from the cloned identities. It will be difficult to obtain ground truth in social networks that do not have any in-built mechanism for verification. Owing to computation limitations, we restrict ourselves to suspected 1,614 clones of the top ten celebrities on Twitter only from India. Therefore, we have a limited and biased dataset. Nevertheless, it is a good first step. In the future, it would be nice to extend the work on celebrities in other countries as well to understand the influence of cultural factors on the clone behaviors. We conveniently selected celebrities as victims because ground truth for them is readily available,

and they have more clones than ordinary persons. Lastly, while the accuracies of behavioral prediction of promotion (86%), indecent (95%) and advisory (92%) are quite decent, at the same time, the accuracies for categories like attention (74%) and opinions (63%) are way too less to be of practical use. More data needs to be collected to improve accuracies for predicting these behaviors. Moving forward, this work can also be extended to build an application that alerts celebrities whenever any clone indulges in any misbehavior. We understand that every celebrity would have a public relations team, who can benefit from such an application.

7 Conclusion

In this work, we present our solution approach for the behavioral characterization of clones. We *recast* the problem as a binary classification problem and conventional classifiers are applied and empirically evaluated. We extract an exhaustive set of features from network, content, and profile of celebrity clone identities. Best classifiers achieve accuracies of 86%, 95%, 74%, 92% and 63% for five clone behaviors namely promotion, indecency, attention seeking, advisory, and opinionated, respectively.

References

1. List of 723 bad words to blacklist & how to use facebook's moderation tool. Front Gate Media (May 2014), https://www.frontgatemedia.com/a-list-of-723-bad-words-to-blacklist-and-how-to-use-facebooks-moderation-tool/, [Online; posted 12-May-2014]
2. What is facebook cloning and how can i protect myself from it? Hoax Slayer (July 2017), https://www.hoax-slayer.net/what-is-facebook-cloning-and-how-can-i-protect-myself-from-it/, [Online; posted 25-July-2017]
3. Aggarwal, A., Kumar, S., Bhargava, K., Kumaraguru, P.: The follower count fallacy: Detecting twitter users with manipulated follower count (2018)
4. Author, C.: Magic marketing words you should be using. Vertical Response (September 2017), https://www.verticalresponse.com/blog/the-30-magic-marketing-words/, [Online; posted 19-September-2017]
5. Bilge, L., Strufe, T., Balzarotti, D., Kirda, E.: All your contacts are belong to us: automated identity theft attacks on social networks. In: Proceedings of the 18th international conference on World wide web. pp. 551–560. ACM (2009)
6. Buxton, M.: The social scam: For a-listers, imposters still loom large. Refinery29 (May 2018), https://www.refinery29.com/2018/05/195519/celebrity-impersonation-accounts-fake-instagram-twitter, [Online; posted 2-May-2018]
7. Cao, Q., Sirivianos, M., Yang, X., Pregueiro, T.: Aiding the detection of fake accounts in large scale social online services. In: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. pp. 15–15. USENIX Association (2012)
8. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., Tesconi, M.: Fame for sale: efficient detection of fake twitter followers. Decision Support Systems 80, 56–71 (2015)

9. Glover, R.: Building a twitter clone. The Meteor Chef (August 2017), https://themeteorchef.com/tutorials/building-a-twitter-clone, [Online; posted 31-August-2017]
10. Goga, O., Venkatadri, G., Gummadi, K.P.: The doppelgänger bot attack: Exploring identity impersonation in online social networks. In: Proceedings of the 2015 Internet Measurement Conference. pp. 141–153. ACM (2015)
11. Gupta, A., Kaushal, R.: Towards detecting fake user accounts in facebook. In: Asia Security and Privacy (ISEASP), 2017 ISEA. pp. 1–6. IEEE (2017)
12. He, B.Z., Chen, C.M., Su, Y.P., Sun, H.M.: A defence scheme against identity theft attack based on multiple social networks. Expert Systems with Applications 41(5), 2345–2352 (2014)
13. Jin, L., Takabi, H., Joshi, J.B.: Towards active detection of identity clone attacks on online social networks. In: Proceedings of the first ACM conference on Data and application security and privacy. pp. 27–38. ACM (2011)
14. Kharaji, M.Y., Rizi, F.S., Khayyambashi, M.R.: A new approach for finding cloned profiles in online social networks. arXiv preprint arXiv:1406.7377 (2014)
15. Lips, A.: Everyone wants to get verified on social media, but it's not usually an easy process. Social Media Week (March 2018), https://socialmediaweek.org/blog/2018/03/can-i-get-verified-verification-guidelines-for-social-media/, [Online; posted 16-March-2018]
16. Parmar, B.: Protecting against spear-phishing. Computer Fraud & Security 2012(1), 8–11 (2012)
17. Sen, I., Aggarwal, A., Mian, S., Singh, S., Kumaraguru, P., Datta, A.: Worth its weight in likes: Towards detecting fake likes on instagram. In: Proceedings of the 10th ACM Conference on Web Science. pp. 205–209. ACM (2018)
18. Slotkin, J.: Twitter 'bots' steal tweeters' identities. Market Place (May 2013), https://www.marketplace.org/2013/05/27/tech/twitter-bots-steal-tweeters-identities, [Online; posted 27-May-2013]
19. Viswanath, B., Mondal, M., Clement, A., Druschel, P., Gummadi, K.P., Mislove, A., Post, A.: Exploring the design space of social network-based sybil defenses. In: Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on. pp. 1–8. IEEE (2012)
20. Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M., Zheng, H., Zhao, B.Y.: Social turing tests: Crowdsourcing sybil detection. arXiv preprint arXiv:1205.3856 (2012)
21. Yu, H., Kaminsky, M., Gibbons, P.B., Flaxman, A.: Sybilguard: defending against sybil attacks via social networks. In: ACM SIGCOMM Computer Communication Review. vol. 36, pp. 267–278. ACM (2006)