DNA-Inspired Online Behavioral Modeling and Its Application to Spambot Detection

Stefano Cresci, *Institute of Informatics and Telematics, IIT-CNR*

Roberto Di Pietro, *Nokia Bell Labs, University of Padova, IIT-CNR*

Marinella Petrocchi, *Institute of Informatics and Telematics, IIT-CNR*

Angelo Spognardi, *DTU Compute, Technical University of Denmark*

Maurizio Tesconi, *Institute of Informatics and Telematics, IIT-CNR*

A novel approach to modeling online user behavior extracts and analyzes digital DNA-inspired sequences from users' online actions. Standard DNA analysis techniques can then discriminate between genuine and spambot accounts.

Social media platforms give Internet users the opportunity to interact and achieve myriads goals, from planning social events to engaging in commercial transactions. Modeling and analyzing online user behaviors deserves attention for a variety of reasons. One is to mine substantial information about events of public interest.[1]

In addition, linking behaviors to a ground truth in the past can help us predict what will likely happen in the future when similar behaviors take place.[2] Furthermore, online behavioral analysis helps detect fictitious and deceitful accounts that could distribute spam or lead to a bias in public opinion.[3]

Current research into online behavior exploits different techniques, such as social interaction graphs,[4,5] textual content,[1,6] and other complex data representations.[2] A unified approach is an open challenge. Our novel notion is to use digital DNA sequences to characterize online user behavior in social media. We believe that digital DNA sequences' high flexibility makes this original modeling technique well suited for different scenarios, with the added potential of opening up new directions of research. By drawing a parallel with biological DNA, we also open up the possibility of drawing on decades of bioinformatics research and development.

We envisage digital DNA sequences assisting in different applications, such as

- letting behavioral patterns emerge from a crowd[2,5] by making use of standard DNA sequence alignment tools and
- defining a behavioral-based taxonomy of online interactions.

For social media, this latter point might mean classifying users as compulsive, curious, lazy, or inactive. Such information could then be

1541-1672/16/$33.00 © 2016 IEEE
Published by the IEEE Computer Society

exploited to launch ad hoc marketing campaigns that convey specific messages to accounts that are more likely to accept such suggestions. This automated approach has the reassuring outcome of avoiding the often frustrating intervention of humans who might not have the means to discriminate patterns by simply inspecting them on an "account-by-account" basis.

Digital DNA

The human genome is the complete set of human genetic information encoded as nucleic-acid sequences. DNA sequences are successions of characters (strings) indicating the order of nucleotides within DNA molecules. The possible characters are A, C, G, and T, representing the four nucleotide bases of a DNA strand: adenine, cytosine, guanine, and thymine. Biological DNA stores the information that directs a living organism's functions and characteristics. DNA sequences are analyzed via bioinformatics techniques, with sequence alignment and motif elicitation being among some of the most well-known approaches for finding sequence commonalities and repetitions. By analyzing common subsequences, it's possible to predict an individual's specific characteristics and uncover relationships among different individuals.

Inspired by biological DNA, we propose modeling online user behavior with strings of characters representing the sequence of a user's online actions. Each action type (such as posting new content or following or replying to a user) can be encoded with a different character, just as in DNA sequences, where characters encode nucleotide bases. According to this paradigm, online user actions would represent the bases of their digital DNA.

Different kinds of user behavior can be observed on the Internet,[6] and digital DNA is a flexible and compact way of modeling such behaviors. The flexibility lies in the possibility of choosing which actions form the sequence. For example, digital DNA sequences on Facebook could include a different base for each user-to-user interaction type: comments (C), likes (L), shares (S), and mentions (M). Then, interactions can be encoded as strings formed by such characters according to the sequence of user-performed actions. Similarly, user-to-item interactions on an e-commerce platform could be modeled by using a base for every product category. User purchasing behaviors could be encoded as a sequence of characters according to the category of products they buy. In this regard, digital DNA shows a major difference from biological DNA, where the four nucleotide bases are fixed; in digital DNA, both the number and the meaning of the bases can change according to the behavior or interaction to be modeled. Just like its biological predecessor, digital DNA is a compact representation of information—for example, a Twitter user's timeline could be encoded as a single string of 3,200 characters (one character per tweet).

In the following, we show our approach's usefulness through its ability to detect Twitter spambots, computer programs that control social accounts with the goal of mimicking real users and sending other users unsolicited messages.[6] Starting from two Twitter datasets with known genuine and spambot accounts, we leverage DNA sequence characterization to let recurrent patterns emerge (groups of spambots share common patterns, as opposed to groups of genuine accounts). We also demonstrate how to apply our methodology to distinguish spambots from genuine accounts in an unknown set of accounts.

Introducing Digital DNA Fingerprinting

Biological DNA fingerprinting is a technique employed by forensic scientists to identify suspects from their DNA. We can use a similar technique for digital DNA fingerprinting to detect spambots on social media. While this is just one possible application for digital DNA, it contributes to understanding how our methodology works in practice.

Academics and platform administrators constantly struggle to keep pace with evolving spambots. New waves of malicious accounts present different and advanced features, making their detection with existing systems extremely challenging.[4,6] This is especially true of the new family of spambots that emerged on Twitter during the last mayoral election in Rome (2014). One of the runners-up used almost 1,000 automated accounts to publicize his policies—surprisingly, these automated accounts were extremely hard to distinguish from genuine ones. Each profile contained detailed personal information and had thousands of genuine followers and friends. Furthermore, they demonstrated tweeting behavior that resembled genuine accounts, with a few tweets posted every day—usually quotes from popular people. The primary anomaly was that every time the candidate posted a new tweet from his official account, all the spambots retweeted it in a time span of just a few minutes. Thus, the candidate reached many more accounts than his own direct followers and managed to alter Twitter's engagement metrics during the electoral campaign.

Here, we focus on two groups of social spambots: the Italian candidate's bot retweeters (Bot1) and bot accounts that spam URLs pointing to several products on Amazon (Bot2). In contrast with classification and supervised approaches, we devise an unsupervised way of detecting spambots that works by comparing their behavior with the aim of finding

similarities among automated accounts. We model the behavior of the two groups of spambots via their digital DNA and compare it to that of a sample of genuine accounts. We exploit digital DNA to study the behavior of groups of users following the intuition that, because of their automated nature, spambots are likely to present higher similarities in their digital DNA with respect to more heterogeneous genuine users.

Mining Groups of Digital DNA Sequences

The basic assumption in our digital DNA fingerprinting technique is that behaviors are considered as sequences of actions; it then characterizes and detects social spambots by grouping similar sequences. Our approach thus falls in the broad field of sequential data mining, yet it presents differences from the tasks commonly performed when working with sequential data. Sequences are ordered lists of symbols from an alphabet B, namely, $s = \{s_1, s_2, ..., s_n\}$, with $s_i \in B$, and are used in information retrieval, part-of-speech tagging, genomics, and time-series analysis, to name but a few. Tasks commonly performed with such sequences are those of sequential supervised learning, time-series prediction, and sequence classification. For instance, in sequential supervised learning, given a sequence $s = \{s_1, s_2, ..., s_n\}$, the goal is to find a sequence of labels $l = \{l_1, l_2, ..., l_n\}$, where each label is associated with an element of the original sequence. Well-known techniques to perform this task are conditional random fields and hidden Markov models. Instead, in time-series prediction, the goal is to find the element s_{t+1}, given a starting sequence $s = \{s_1, s_2, ..., s_t\}$. This goal could be achieved by employing a model taken from the broad family of statistical (auto)regressive techniques. Finally, sequence classification is concerned with assigning a single label l to a whole sequence $s = \{s_1, s_2, ..., s_n\}$ following a supervised approach.

In contrast, our goal here is to analyze a set of unlabeled sequences to find commonalities and differences among them, exploiting the sequential information related to user actions that's captured in digital DNA sequences. In this regard, our task resembles that of unsupervised sequence clustering. However, it further differs from traditional machine learning clustering because we exploit sequential data. Indeed, our sequences are ordered data representations, whereas feature vectors exploited in traditional clustering are unordered (features aren't ordered in feature vectors, and the ordering isn't exploited in the analysis). Moreover, our sequences are variable-length vectors of symbols drawn from a limited alphabet (strings), rather than fixed-length numeric vectors. Thus, traditional distance and similarity metrics for numeric vectors (such as those typically used in clustering) aren't applicable in our case. Working with ordered and variable-length data representations also marks a difference with other techniques recently used for group behavioral analysis, such as those based on hashing[7] and graph mining.[5]

Digital DNA Fingerprinting for Spambot Detection

Just as for human DNA, which can be collected from different places, a user's digital fingerprint can be retrieved from different online contexts. Like the human version, it's sequenced to be compared with other subjects. However, there's a big difference between human and digital DNA: in the former, the bases are four and fixed, whereas in the latter, we can decide both number and type of bases, depending on the type of analysis to be conducted. To show this flexibility, we describe in the following a complete example, in which we use the type of tweets to characterize and analyze the digital DNA of Twitter users.

Characterizing Spambot Behavior

The process of digital DNA fingerprinting has four main steps: behavioral data acquisition, DNA sequence extraction, DNA sequence comparison, and evaluation. To ensure a rigorous evaluation of our detection technique, we created datasets of verified spambots and genuine Twitter accounts. All the accounts in our datasets underwent manual verification—specifically, we certified 50.05 percent (991 accounts) of Bot1 accounts and 89.29 percent (464 accounts) of Bot2 accounts as spambots. Then, we built a dataset of human accounts by randomly contacting Twitter users and asking them simple questions in natural language. We certified the 3,474 accounts that answered as human. For the 4,929 accounts in our datasets, we then collected behavioral data by crawling the content of their Twitter timelines.

To extract digital DNA—that is, to associate each account to a string that encodes its behavioral information—we can use different techniques, depending on the kind of information to be modeled. Similar to feature engineering in machine learning, catching the right granularity for modeling behavior might lead to better detection results, so in our experiments, we modeled account behavior in two ways and evaluated which better captures the nature of automated accounts. In the first experiment, we considered the types of tweets shared (tweet type DNA) and encoded every tweet in the account's timeline with a different character: A for a simple tweet, T for a reply, and C for a retweet. In the second experiment, we considered the content of tweets (tweet content DNA) and designed base A for tweets with URLs, T for tweets with hashtags, C for tweets

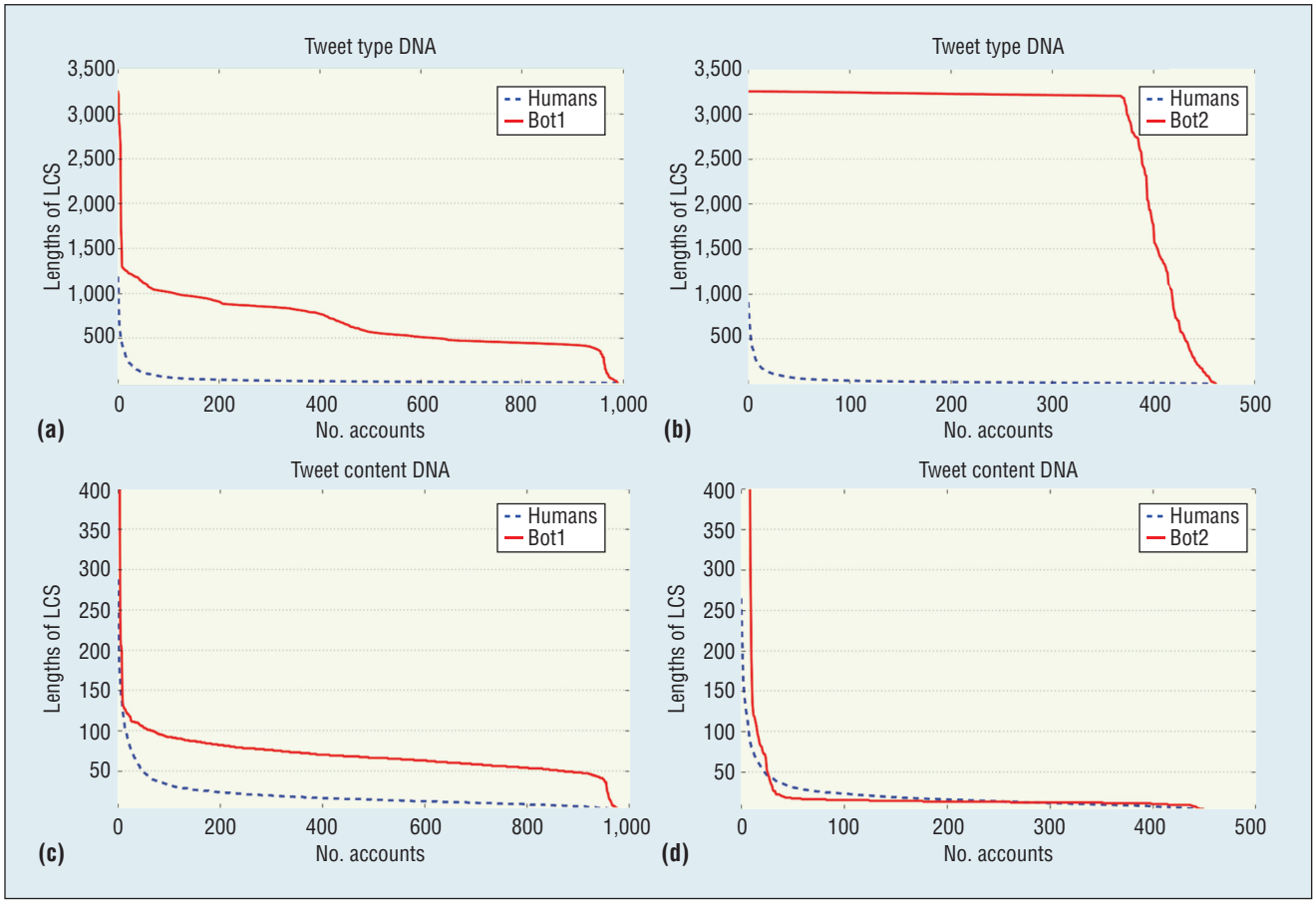


Figure 1. Comparison of digital DNA for different groups of Twitter accounts: (a) tweet type DNA similarity between political candidate retweeters and genuine accounts, (b) tweet type DNA similarity between Amazon product spammers and genuine accounts, (c) tweet content DNA similarity between political candidate retweeters and genuine accounts, and (d) tweet content DNA similarity between Amazon product spammers and genuine accounts.

with mentions, G for tweets with media (pictures, videos), X for tweets with a combination of the previous entities, and N for tweets with none of them (plaintext). Because each Twitter timeline is made up of at most 3,200 tweets, we obtained digital DNA sequences counting up to 3,200 characters.

We consider similarity as a proxy for automation, and, thus, an exceptionally high level of similarity among a large group of accounts serves as a red flag for anomalous behaviors. We quantify similarity by looking at the longest common substring (LCS) among digital DNA sequences. We rely on a linear time algorithm[8] that's based on the generalized suffix tree. Given m strings, the algorithm is able to find, for all $2 \le k \le m$, the LCS to at least k strings. Then, given m DNA sequences and the length of the longest substring common to k of them, we can derive which accounts have such longest substrings in common (which are those k accounts). As the number k of accounts grows, the length of the LCS to all of them shortens. Thus, we're more likely to find a long LCS among a few accounts rather than among large groups. The rationale behind our analysis is that if the LCS is long when k grows, then the accounts that share that LCS have a suspiciously similar behavior.

To verify our claim, we computed the LCS in the datasets of certified spambots and genuine accounts for every k. To be statistically sound, we compared groups of the same size by randomly undersampling genuine accounts to match the number of spambots. Figures 1a and 1b show results of the comparison of tweet type DNA sequences, while Figures 1c and 1d show the results for tweet content DNA sequences. All plots of Figure 1 show the lengths of LCSs for all possible group sizes: for example, in Figure 1a, there exists a group of 400 accounts in Bot1 with a substring of length 774 in common. From Figures 1a and 1b, it's clear that the LCS of both groups of spambots is rather long even when the number of accounts grows. This is strikingly evident for Bot2 (see Figure 1b). In contrast, genuine accounts show little to no similarity. For both the spambot groups, we observe a sudden drop in LCS length when k gets close to the

group size. The human group has an exponential decay and rapidly reaches the smallest values of the LCS length.

Figures 1c and 1d are related to tweet content DNA sequences, and results are noticeably different from those obtained with tweet type DNA. All the LCS curves have much lower values, but the biggest difference between tweet content DNA and tweet type DNA is for Bot2. Although tweet type DNA models almost perfectly the high similarity of Bot2 accounts, tweet content DNA is unable to show a significant difference between genuine and bot accounts. These results suggest that modeling the behavior of Twitter accounts according to tweet type, rather than content, is more effective for discriminating Bot2 spambots.

Uncovering Novel Twitter Spambots

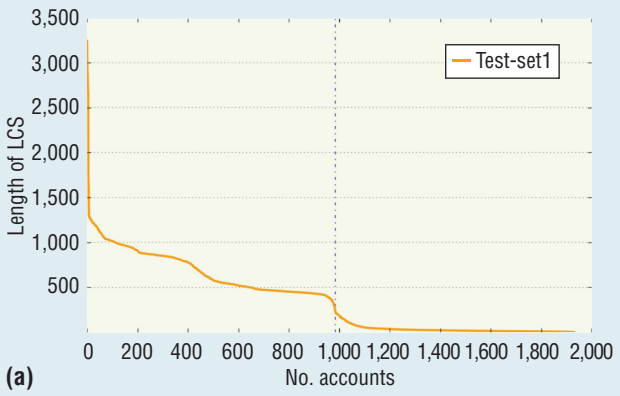
As shown earlier, properly extracted digital DNA sequences provide evidence of similar activities in large groups, which can lead to suspicions of a high degree of automation for those accounts behaving in a similar way. Building on the promising results obtained with tweet type DNA, we can exploit the characteristics of LCS curves to effectively detect groups of Twitter spambots. Our detection technique can be applied to a group of unlabeled accounts to check whether those with a suspiciously similar behavior are present. In a group with mixed bot and genuine accounts, almost only the bot accounts will have long DNA substrings in common. Thus, by identifying the group of accounts that share a long LCS, we can obtain a set of suspicious accounts.

Henceforth, Test-set1 and Test-set2 refer to unlabeled groups where genuine accounts have been mixed with the ones from Bot1 and Bot2, respectively. The plots in Figures 2a and 2b show the lengths of LCSs for all the possible values of k accounts in Test-set1 and Test-set2. In these plots, we looked for the points where the LCS curves exhibit a sudden drop to a very low value, because those points might represent thresholds between groups of similar accounts. Observing Figure 2a, we can see a continuous decrease in LCS length as the number of accounts grows and a sudden drop just before reaching 1,000 accounts. The drop is even more evident in Figure 2b, where there's a steep fall just before 400 accounts. LCS curves in both plots keep approaching 1 as the number of accounts grows. The steep drops in LCS curves highlight areas where LCS length remains practically unchanged even for significantly different numbers of accounts considered. In fact, such plateaus in LCS curves are strictly related to homogeneous groups of highly similar accounts.

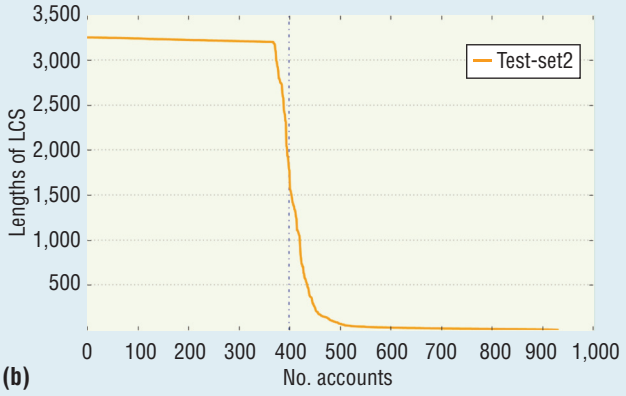
Note that it's possible to observe multiple plateaus in a single LCS curve, as in the case of Figure 2a. This represents a situation where multiple (sub)groups exist among the whole set of considered accounts. Furthermore, the steeper and the more pronounced the drop in the LCS curve, the more different are the two groups of accounts split by that drop.

Building on these interesting characteristics, we devised a methodology to exploit drops in LCS curves to identify groups of similar accounts. Specifically, we exploit the derivative of LCS curves ΔLCS/Δaccounts to highlight the points corresponding to the steepest drops. Such points, appearing as negative peaks in the derivative plot, represent good candidate splitting points to detect groups of dissimilar accounts. Note that it's possible to obtain several candidate splitting points, which might also be ranked by derivative value (steepness of the corresponding drop). Then, a hierarchical top-down (divisive) approach can be applied by repeatedly splitting the set of accounts based on the ranked candidate points, for instance, when the LCS curve exhibits multiple plateaus and steep drops, leading to a dendrogram-like structure. From the LCS curves in Figures 2a and 2b, we identified the best splitting points and assumed that the suspiciously similar accounts shared the LCS just before such splitting points, namely, the 983 accounts with an LCS around 400 for Test-set1 and the 400 accounts with an LCS of around 1,750 for Test-set2.

Finally, we evaluated the effectiveness of the DNA-based detection technique by considering the evaluation metrics precision, recall, specificity, accuracy, F-measure, and the Matthews correlation coefficient (MCC). To thoroughly evaluate the DNA fingerprinting technique, we compared our detection results with those obtained from several other state-of-the-art approaches, namely, the supervised one by Chao Yang and colleagues[4] and the unsupervised approaches by Zachary Miller and colleagues[9] and Faraz Ahmed and Muhammad Abulaish.[10] Yang's work[4] provides a machine learning classifier that infers whether a Twitter account is genuine or a spambot by relying on account relationships, tweeting timing, and level of automation. We reproduced these classifiers because the authors kindly provided us with their training sets. In the other works,[9,10] the authors define a set of machine learning features and apply clustering algorithms. Specifically, Miller[9] proposed modified versions of the DenStream and StreamKM++ algorithms (respectively based on DBSCAN and k-means) and applied them to the detection of spambots over the Twitter stream, and Ahmed and Abulaish[10] exploited the Euclidean distance between feature vectors to build a similarity graph of accounts. Graph clustering and community detection



(a)



(b)

Detection system	Detection results				Evaluation metrics					
	TP	*TN*	*FP*	*FN*	*Precision*	*Recall*	*Specificity*	*Accuracy*	*F-measure*	*MCC*
C. Yang and colleagues[4]	169	811	131	822	0.563	0.170	0.860	0.506	0.261	0.043
Z. Miller and colleagues[9]	355	657	285	636	0.555	0.358	0.698	0.526	0.435	0.059
F. Ahmed and colleagues[10]	935	888	54	56	0.945	0.944	0.945	0.943	0.944	0.886
DNA fingerprinting	963	924	18	28	0.982	0.972	0.981	0.976	0.977	0.952

(c)

Detection system	Detection results				Evaluation metrics					
	TP	*TN*	*FP*	*FN*	*Precision*	*Recall*	*Specificity*	*Accuracy*	*F-measure*	*MCC*
C. Yang and colleagues[4]	190	397	71	274	0.727	0.409	0.848	0.629	0.524	0.287
Z. Miller and colleagues[9]	142	306	162	322	0.467	0.306	0.654	0.481	0.370	–0.043
F. Ahmed and colleagues[10]	428	427	41	30	0.913	0.935	0.912	0.923	0.923	0.847
DNA fingerprinting	398	468	0	66	1.000	0.858	1.000	0.929	0.923	0.867

(d)

Figure 2. DNA similarity of the two test-sets and evaluation of spambot detection results. (a) DNA similarity of a mixed group of genuine accounts and political candidate retweeters (Test-set1). (b) DNA similarity of a mixed group of genuine accounts and Amazon product spammers (Test-set2). (c) Evaluation of spambot detection results for a mixed group of genuine accounts and political candidate retweeters (Test-set1). (d) Evaluation of spambot detection results for a mixed group of genuine accounts and Amazon product spammers (Test-set2). According to the feature set of Faraz Ahmed and colleagues,[10] a few accounts had null values for all features, resulting in the impossibility of applying the clustering algorithm to such accounts.

algorithms are then applied to identify groups with similar accounts in the graph.

Figures 2c and 2d show the results of this comparison. Notably, the DNA fingerprinting detection technique outperforms all other approaches, achieving MCC = 0.952 for Test-set1 and MCC = 0.867 for Test-set2. There's a clear performance gap between other approaches[4,9] with respect to our proposed approach and that of Ahmed and Abulaish.[10] Yang's[4] supervised approach proved unable to accurately distinguish spambots from genuine accounts, as demonstrated by the considerable number of false negatives and the resulting very low recall. This result supports our initial claim that this new wave of bots is surprisingly similar to genuine accounts: they're exceptionally hard to detect if considered individually. Moreover, among the 126 features Miller proposed,[9] 95 are based on the textual content of tweets. However, novel social spambots tweet content similar to that of genuine accounts (retweets of genu-

THE AUTHORS

Stefano Cresci is a PhD student in the Information Engineering Department at the University of Pisa and a research fellow at IIT-CNR in Pisa, Italy. His research interests include social media mining and knowledge discovery. He's a student member of IEEE and a member of the IEEE Computer Society. Contact him at stefano.cresci@iit.cnr.it.

Roberto Di Pietro is Security Research Global Head at Nokia Bell Labs, Paris. His research interests include security, privacy, distributed systems, computer forensics, and analytics. Roberto is also affiliated with the University of Padua and IIT-CNR in Pisa, Italy. Contact him at roberto.di_pietro@nokia-bell-labs.com.

Marinella Petrocchi is a researcher at IIT-CNR in Pisa, Italy. Her research interests include data privacy, personalization on the Internet, data quality, and trustworthiness. Contact her at marinella.petrocchi@iit.cnr.it.

Angelo Spognardi is an assistant professor at DTU Compute, Denmark. His research interests include social networks modeling and analysis, cyber-physical systems and network security, and privacy. Contact him at angsp@dtu.dk.

Maurizio Tesconi is a researcher at IIT-CNR in Pisa, Italy. His research interests include social Web mining, social network analysis and visual analytics within the context of Open Source Intelligence. He's also part of the permanent team of the European Laboratory on Big Data Analytics and Social Mining, performing advanced research and analyses on the emerging challenges posed by big data. Contact him at maurizio.tesconi@iit.cnr.it.

ine tweets and famous quotes). Ahmed and Abulaish's approach,[10] however, proved effective in detecting our considered spambots, showing an MCC = 0.886 for Test-set1 and MCC = 0.847 for Test-set2. Employing only seven features, Ahmed and Abulaish[10] focus on retweets, hashtags, mentions, and URLs, thus analyzing the accounts along the dimensions exploited by spammers. However, although it achieved an overall good performance on the spambots, this approach[10] could lack reusability across other groups of spambots with different behaviors, such as those perpetrating a follower fraud.[5,11] In contrast, our DNA-inspired technique is flexible enough to highlight suspicious similarities among groups of accounts without focusing on specific characteristics.

Finally, it's worth noting that some state-of-the-art approaches for spambot detection require a large number of data-demanding features. For instance, approaches based on graph mining are more demanding in terms of the data required to perform the detection.[11] Our DNA fingerprinting technique exploits only Twitter timeline data to perform spambot detection. Furthermore, it doesn't require a training phase and can be employed pretty much like a clustering algorithm, in an unsupervised fashion.

We envision a possibility of exploiting the results of our DNA-inspired technique as a feature in a more complex detection system. Indeed, different types of DNA (such as tweet type and content DNA) can be exploited to model user behavior in different directions. The results of these models could be used simultaneously in an ensemble or voting system. The already interesting results achieved by exploiting only one type of digital DNA—namely, tweet type DNA—could represent promising ground for further experimentation and research. ■

Acknowledgments

We thank the anonymous reviewers for their fruitful comments and suggestions. This research has been partially supported by Registro.it under the MIB, My Information Bubble, project. This research is also supported in part by the SoBigData Research Infrastructure project, funded by the European Commission under the H2020 programme, and by the Tuscany Italian regional project SmartNews.

References

1. A. Tsakalidis et al., "Predicting Elections for Multiple Countries Using Twitter and Polls," *IEEE Intelligent Systems*, vol. 30, no. 2, 2015, pp. 10–17.
2. K. Li and Y. Fu, "Prediction of Human Activity by Discovering Temporal Sequence Patterns," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, 2014, pp. 1644–1657.
3. H. Liu, J. Han, and H. Motoda, "Uncovering Deception in Social Media," *Social Network Analysis and Mining*, vol. 4, no. 1, 2014, p. 162.
4. C. Yang, R. Harkreader, and G. Gu, "Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers," *IEEE Trans. Information Forensics and Security*, vol. 8, no. 8, 2013, pp. 1280–1293.
5. M. Jiang et al., "Catching Synchronized Behaviors in Large Networks: A Graph Mining Approach," *ACM Trans. Knowledge Discovery from Data*, vol. 10, no. 4, 2016, pp. 1–27.
6. X. Hu, J. Tang, and H. Liu, "Online Social Spammer Detection," *Proc. AAAI Conf. Artificial Intelligence*, 2014, pp. 59–65.
7. M. Ou et al., "Probabilistic Attributed Hashing," *Proc. AAAI Conf. Artificial Intelligence*, 2015, pp. 2894–2900.
8. M. Arnold and E. Ohlebusch, "Linear Time Algorithms for Generalizations of the Longest Common Substring Problem," *Algorithmica*, vol. 60, no. 4, 2011, pp. 806–818.
9. Z. Miller et al., "Twitter Spammer Detection Using Data Stream Clustering," *Information Sciences*, vol. 260, 2014, pp. 64–73.
10. F. Ahmed and M. Abulaish, "A Generic Statistical Approach for Spam Detection in Online Social Networks," *Computer Comm.*, vol. 36, no. 10, 2013, pp. 1120–1129.
11. S. Cresci et al., "Fame for Sale: Efficient Detection of Fake Twitter Followers," *Decision Support Systems*, vol. 80, 2015, pp. 56–71.