SOCIAL FINGERPRINTING: DETECTION OF SPAMBOT GROUPS THROUGH DNA-INSPIRED BEHAVIORAL MODELING

Seminar Report

Submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

of

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Submitted By

CLINT MATHEWS

Department of Computer Science & Engineering
Mar Athanasius College Of Engineering Kothamangalam

SOCIAL FINGERPRINTING: DETECTION OF SPAMBOT GROUPS THROUGH DNA-INSPIRED BEHAVIORAL MODELING

Seminar Report

Submitted in partial fulfillment of the requirements for
the award of degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

of

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Submitted By

CLINT MATHEWS

Department of Computer Science & Engineering
Mar Athanasius College Of Engineering Kothamangalam

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MAR ATHANASIUS COLLEGE OF ENGINEERING
KOTHAMANGALAM

CERTIFICATE

This is to certify that the report entitled **Social Fingerprinting: Detection of spambot groups through DNA-inspired behavioral modeling** *submitted by* ***Mr. CLINT MATHEWS****, Reg. No.* ***MAC15CS023*** *towards partial fulfillment of the requirement for the award of Degree of Bachelor of Technology in Computer science and Engineering from APJ Abdul Kalam Technological University for December 2018 is a bonafide record of the seminar carried out by him under our supervision and guidance.*

....	....	....
Prof. Joby George	**Prof. Neethu Subash**	**Dr. Surekha Mariam Varghese**
Faculty Guide	***Faculty Guide***	***Head of the Department***

Date: Dept. Seal

ACKNOWLEDGEMENT

First and foremost, I sincerely thank the God Almighty for his grace for the successful and timely completion of the seminar.

I express my sincere gratitude and thanks to Dr. Solly George, Principal and Dr. Surekha Mariam Varghese, Head Of the Department for providing the necessary facilities and their encouragement and support.

I owe special thanks to the staff-in-charge Prof. Joby george, Prof. Neethu Subash and Prof. Joby Anu Mathew for their corrections, suggestions and sincere efforts to co-ordinate the seminar under a tight schedule.

I express my sincere thanks to staff members in the Department of Computer Science and Engineering who have taken sincere efforts in helping me to conduct this seminar.

Finally, I would like to acknowledge the heartfelt efforts, comments, criticisms, co-operation and tremendous support given to me by my dear friends during the preparation of the seminar and also during the presentation without whose support this work would have been all the more difficult to accomplish.

ABSTRACT

Spambot detection in online social networks (OSN) is a long-lasting challenge. A new wave of social spambots has emerged, with advanced human-like characteristics which goes undetected even by current state-of-the-art algorithms. Efficient spambots detection can be achieved by analyzing collective behaviors. Digital DNA technique models behaviors of OSN users. Digital account is encoded in a sequence of characters. Measuring similarity for such digital DNA sequences. Similarity between groups of users characterizes both genuine accounts and spambots. Using such characterization the Social Fingerprinting technique is able to discriminate among spambots and genuine accounts in both a supervised and an unsupervised fashion. Effectiveness of Social Fingerprinting is evaluated by comparing it with existing algorithms. In particular social fingerprinting has the possibility of using DNA analysis techniques to study online users behaviors and can rely on a limited number of account characteristics.

Contents

List of Figures

Figure No. Name of Figures Page No.

List of Abbreviations

OSN	Open Social Networks
DNA	Deoxyribonucleic Acid
LCS	Longest Common Substring
ROC	Receiver Operating Characteristic
MCC	Matthews Correlation Coefficient
DBSCAN	Density-based spatial clustering of applications with noise
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

Introduction

Online social networks (OSNs) provide Internet users with the opportunity to discuss, get informed, express themselves, and interact for achieving a lot of goals, such as planning events and engaging in commercial transactions .In a word, users rely on online services to say to the world what they are, think, do; and viceversa, they learn the same about the other subscribers.

Quite naturally, the widespread availability and ease of use have made OSNs the ideal setting for the proliferation of fictitious and malicious accounts. While hiding a real identity is sometimes motivated by the harmless side of ones personality, there exist however deceitful situations where accounts of social platforms are created and managed to distribute unsolicited spam, advertise events and products of doubtful legality, sponsor public characters and, ultimately, lead to a bias within the public opinion.

Peculiarity of social spambots is that they evolve over time, adopting sophisticated techniques to evade early established detection approaches, such as those based on textual content of shared messages, posting patterns ,and social relationships. As evolving spammers became clever in escaping detection, for instance by changing discussion topics and posting activities, researchers stayed in line with the times and proposed complex models based on the interaction graphs of the accounts under investigation.

Standard classification approach, where the single account is evaluated according to a set of established features tested over known datasets, is no longer successful. Instead, the intuition is that the key factor for spotting new social spambots is focusing on the collective behavior of groups of accounts, rather then on single behaviors.

Online behavioral modeling: Behaviors are modeled via digital DNA, namely strings of characters, each of them encoding one action of the online account under investigation. Similarly to biological DNA, digital DNA allows a compact representation of information.

Digital DNA is a flexible model, able to represent different actions, on different social platforms, at different levels of granularity. We extract and analyze digital DNA sequences from the behaviors of OSNs users, and we use Twitter as a benchmark to validate our proposal. We obtain a compact and effective DNA-inspired characterization of user actions. Then, we apply standard DNA analysis techniques to discriminate between genuine and spambot accounts on Twitter.

Spambot detection: Groups of spambots share common patterns, as opposite to groups of genuine accounts. As a concrete application of this outcome, we apply our Social Fingerprinting methodology to tell apart spambots from genuine accounts, within an unknown set of accounts. The excellent performances obtained in terms of standard classifiers-based indicators like F-Measure, Accuracy, Precision, and Recall which supports the quality and viability of the Social Fingerprinting technique.

Twitter spambot detection is a specific use case on a specific social network, our proposed Social Fingerprinting technique is platform and technology agnostic, hence paving the way for diverse behavioral characterization tasks. Indeed, we believe that the high flexibility and applicability of digital DNA sequences make this new modeling approach suitable to represent different scenarios, with the potential to open new directions of research. Making use of standard DNA sequences alignment tools, our approach has the comfortable outcome of avoiding the often frustrating intervention of humans, who may not have the means to discriminate patterns by simply inspecting on an account by account basis.

The remainder of this work is organized as follows. The next section presents a survey of relevant work in the field of social networks spambot detection. Following that, we introduce the notion of digital DNA and propose a similarity measure for digital DNA sequences. Then the characteristics of online accounts based on their DNA sequences and the results of our approach for Twitter spambot detection. The final section draws a conclusion.

Related works

During a period spanning the last six years, the academic literature has seen the flowering of scientific approaches to model and analyze anomalous accounts on social networks. In particular, Twitter has gained a lot of attention, since the platform massively features different kinds of peculiar subscribers, such as spammers, bots, cyborgs, and fake followers.Spammers are those accounts that advertise un-solicited and often harmful content, containing links to malicious pages[1], bots are computer programs that control social accounts, as stealthy as to mimic real users[1] , while cyborgs interweave characteristics of both manual and automated behavior[1]. Finally, there are fake followers, namely accounts massively created to follow a target account and that can be bought from online markets, also attracting the interest of mass media.

2.1 Established techniques

As an example for spam detection, a branch of research mined the textual content of tweets, others studied the redirection of embedded URLs in tweets or classified the URLs landing pages. Work also moved beyond the difficulty of labeling those tweets without URLs as spam tweets, by proposing a composite tool, able to match incoming tweets with underlying templates commonly used by spammers.

Other work investigated spammers through a multi-feature approach, including features on the profile, the behavior, and the time line of an account.Considering fake Twitter followers since both spammers, bots, and genuine accounts could fall in this category applying existing series of rules and features on humans and fake followers.The main contributions were pruning those rules and features that behaved worst in detecting fake followers and implementing a classifier which significantly reduces over-fitting and cost for data gathering. Majority of those contributions were grounded on the assumption that it is possible to recognize an account as genuine or not, based on a series of characteristics featured by that account. The classification takes place on the single account and was found with anomalies.

2.2 Emerging trends

New social bots are rising, whose peculiarity emerges only when considering their collective behavior. If the accounts are considered one by one, they are no more distinguishable from genuine ones. We claim that such social spambots represent the third and most novel generation of spambots. Here we show how digital DNA represents a powerful basis for the detection of the third generation of spambots.Furthermore, they propose an algorithm to spot users featuring unexpected behaviors. If a collective online action happens once, then that action is not necessarily fraudulent. Instead, if that collective action repeats over time, especially in reaction to the same kind of event, it probably represents an anomalous activity. In particular, the work focuses on retweeting activities, defines features for retweet threads characterization, and proposes a methodology for catching synchronized frauds.

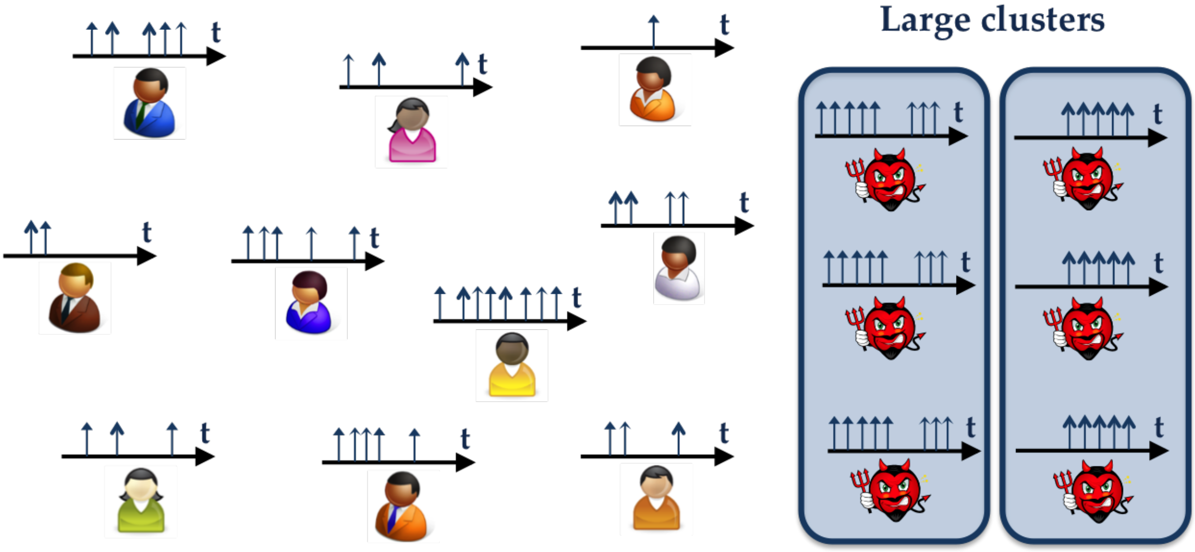


Fig. 2.1: Synchrotrap detection

SynchroTrap aims at detecting loosely synchronized behaviors for a broad range of social network applications. Fig 2.1 shows how the method works at a particular time period. Time is an important dimension for SynchroTrap , since the methodology forms clusters on the basis of equal actions performed by online accounts within the same time interval.

2.3 Comparison

The main differences of this approach with respect to the existing one. Firstly consider a single dimension as the basis to let groups of social accounts emerge the digital DNA, the sequence of characters encoding the accounts behavior. Secondly, properties of the social graph are not considered which leads to the significant advantage of reducing the cost for data gathering. Approaches that are based on graph mining generally rely on a large quantity of data and can require computationally expensive algorithms to perform their detection. Only Twitter timeline data is exploited to perform spambots detection. Third, the analysts can leverage powerful set of tools developed over decades for DNA analysis to validate their working hypotheses on online spambots behaviors. Furthermore,the DNA-inspired modeling focuses on the concept of sequence, namely ordered lists of symbols, with variable length, and taken from a relatively small alphabet. This marks a clear separation from other well known behavioral analysis techniques that do not consider the ordering of the elements, like hashing . As discussed later on, the results are promising and they lead us to believe that digital DNA is a simple and compact, yet powerful, mean to detect the novel waves of social spambots.

Proposed method

3.1 Twitter datasets

The different Twitter datasets that constitute the real-world data used in experiments. Specifically, the collected months worth of data about the activities of a random sample of genuine humanoperated accounts and of two different families of spambots.

3.1.1 Bot1 and bot 2 datasets

A first dataset of spambots was created after observing the activities of a novel group of social bots that we discovered on Twitter during the last Mayoral election in Rome,in 2014. One of the runners-up employed a social media marketing firm for his electoral campaign, that made use of almost 1,000 automated accounts on Twitter to publicize his policies. Surprisingly, automated accounts were found to be similar with genuine ones in every way. Every profile was accurately filled with detailed yet fake personal information such as a stolen photo, short-bio, location, etc. Those accounts also represented credible sources of information since they all had thousands of followers and friends, the majority of which were genuine users. Furthermore, the accounts showed a tweeting behavior which was apparently similar to those of genuine accounts, with a few tweets posted every day, mainly quotes from popular people. However, every time the political candidate posted a new tweet from his official account, all the automated accounts retweeted it in a time span of just a few minutes. By resorting to this farm of bot accounts, the political candidate was able to reach many more genuine accounts in addition to his direct followers and managed to alter Twitter engagement metrics during the electoral campaign.

Second group of social bots whose intent was to advertise a subset of products on sale on the Amazon.com e-commerce platform. This time the deceitful activity was carried out by spamming URLs pointing to the advertised products. However, similarly to the retweeters of the Italian political candidate, also this family of spambots interleaved spam tweets with many harmless and genuine ones. Henceforth, we refer to the spambots retweeters of the Italian political candidate as Bot1 and to those spambots advertising Amazon.com products as Bot2.

3.1.2 Comparison

As for a mere comparison of the Twitter profiles, it is nearly impossible to tell apart the spambots from the genuine account. Worryingly, this is the same scenario that Twitter users are typically presented to, while browsing the social platform. To make the situation even worse the novel social spambots also employ social engineering techniques, such as the profile picture of a young attractive woman and the occasional posting of provocative tweets, in order to lure genuine accounts. As such, any threat spread out by social spambots are more likely to result in a successful attack with respect to those spread by traditional spambots.

After identifying possible spambots, we exploited a Twitter crawler to collect data about all the accounts we suspected to belong to the two groups of spambots. All the accounts collected in this process have then undergone a manual verification phase to certify their automated nature. Specifically, the spambots of our datasets were annotated by two tech-savvy post-graduate students, with yearly experience on Twitter and social media. To evaluate the interannotator agreement, the well-known Cohens Kappa evaluation metric was used and was excellent for both groups. The disagreements between the two annotators have been resolved by a superannotator, a Ph.D. student with yearly experience in sybil and spambot detection. Summarizing, among all the distinct retweeters of the Italian political candidate, 50.05% were certified as spambots. Similarly, 89.29% of the accounts that tweeted suspicious Amazon. com URLs were also certified as spambots. These 2 sets of accounts represent our ground truth of social spambots. Then, in order to build a dataset of certified human accounts, we randomly contacted Twitter users by asking them simple questions in natural language, following a hybrid crowdsensing approach . Replies to our questions were manually verified .

Cohen's kappa coefficient (k) is a statistic which measures inter-rater agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, as takes into account the possibility of the agreement occurring by chance. There is controversy surrounding Cohens kappa due to the difficulty in interpreting indices of agreement. Some researchers have suggested that it is conceptually simpler to evaluate disagreement between items.[2]

For all the accounts of our datasets, we then collected behavioral data by crawling the content of their Twitter pages. Furthermore, we also collected data about all their direct followers and friends, and about all the accounts they interacted with in their tweets.

3.2 Digital deoxyribonucleic acid

The human genome is the complete set of genetic information on humans and it is encoded in the form of nucleic acid DNA sequences. A DNA sequence is a succession of characters, a string that indicates the order of nucleotides within a DNA molecule. The possible characters are A, C, G, and T, representing the four nucleotide bases of a DNA strand: adenine, cytosine, guanine, thymine. Biological DNA stores the information which directs functions and characteristics of a living organism.

Nowadays, DNA sequences are exploited worldwide. DNA sequences can be read from raw biological material through DNA sequencing methods. Currently, such sequences are stored in sequence databases and they are analyzed by means of bioinformatics techniques. Among the most well known and widely adopted analysis techniques are sequence alignment and repetition/motif elicitation. One of the main goals of these techniques is to find commonalities and repetitions among DNA sequences. Indeed, via an analysis of common subsequences and substrings it is possible to predict specific characteristics of the individual and to uncover relationships between different individuals. By drawing a parallel with biological DNA, we envisage the possibility to model OSNs users behaviors and interactions by means of strings of characters, representing the sequence of their actions. Online actions such as posting new //content, replying to another user, following an account can be encoded with different characters, similarly to DNA sequences, where the A, C, G, T characters encode the four nucleotide bases. According to this parallelism, a users actions represent the bases of his/her digital DNA[3]. There exist different kinds of user behaviors on OSNs. Digital DNA is a flexible and compact yet effective way of modeling such behaviors. Its flexibility lies in the possibility to choose which actions to consider while building the DNA sequence. For example, a digital DNA sequence can be built to model user-to-user interactions on Facebook by defining a different base for every possible interaction type, such as comments (base C), likes (base L), shares (base S) and mentions (base M). Users interactions can then be encoded as strings composed of the C, L, S and M characters according to the sequence of actions they perform. Similarly, it is possible to model users tweeting behaviors on Twitter by defining different bases for tweets, retweets, and replies. Users tweeting behaviors can then be encoded as a sequence of characters according to the sequence of tweets they post. To this regard, digital DNA shows a major difference with 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/interaction one aims to model. Similarly to its biological counterpart, digital DNA is also a compact representation of information. For example, the time line of a Twitter user can be encoded as a single string of 3,200 characters,one character per tweet.

There is a vast number of algorithms and techniques to draw upon for the analysis of digital DNA sequences. Indeed, many of the techniques developed in the last few years in the field of bioinformatics for the analysis of biological DNA can be leveraged to study the characteristics of digital DNA as well. In the following we give a general definition of digital DNA, and we introduce the Twitters digital DNA concept, with one of its possible application is spambots detection.

3.2.1 Definition of digital deoxyribonucleic acid sequences and its application on twitter

The bases used to create a digital DNA sequence are represented as a finite set of unique symbols or characters.The representation can also contain all the actions ranging to repeated actions of the user. A digital DNA sequence is an ordered tuple or row vector, of characters, a string whose possible values are defined by the bases of its alphabet.The bases taken into consideration can be used for the string formation and can be used for producing a characteristic string of users behaviour. A limited number of bases in an alphabet can be used to create sequences of arbitrary length.

A sequence s is defined as

$$s = (b_1, b_2, ..., b_n)b_i \in B \forall i = 1, 2, ..., n. \qquad \text{(Equ:3.1)}$$

$b_1, b_2, ..., b_n$ represents bases and B contains all the bases taken into consideration in the digital DNA.A digital DNA sequence based on the B_3 type alphabet can then be obtained by scanning the tweets produced by a user on Twitter and by assigning the T character to every retweet, the C character to every reply, and the A character to every other tweet, in the same order of the tweets generated by the user. An excerpt of a digital DNA sequence generated with the alphabet

$$B_T^3, s = (A; A; A; C; A; T; C; A; A; C; :::). \qquad \text{(Equ:3.2)}$$

Where s represents a digital DNA sequence. A digital DNA sequence can also be represented with a more compact notation as a string instead of a row vector.

$$s = AAACATCAAC :::, \qquad \text{(Equ:3.3)}$$

Another possible way of modeling the content of tweets could have involved the detection of the topic of a tweet. Then, it would have been possible to define an alphabet so as to have a different base for each of the main topics, such as politics, sports, technology, music, etc. Anyway, for the sake of simplicity, in our work we only exploited Twitter entities in order to obtain DNA sequences based on the content of tweets. In the above notations, alphabets are characterized by a subscript example type that identifies the kind of information captured by the bases, and by a superscript that denotes the number N of bases in the alphabet. These two indices are typically enough to unequivocally identify an alphabet.

3.2.2 Longest Common Substring: a similarity measure for digital Deoxyribonucleic acid sequences

$$\mathbb{B}^{3}_{content} = \left\{ \begin{array}{l} \mathtt{N} \longmapsfrom \text{tweet contains no entities (plain text),} \\ \mathtt{E} \longmapsfrom \text{tweet contains entities of one type,} \\ \mathtt{X} \longmapsfrom \text{tweet contains entities of mixed types} \end{array} \right\} = \{\mathtt{N}, \mathtt{E}, \mathtt{X}\}$$

$$\begin{aligned} \mathbb{B}^{6}_{content} &= \left\{ \begin{array}{l} \mathtt{N} \longmapsfrom \text{tweet contains no entities (plain text),} \\ \mathtt{U} \longmapsfrom \text{tweet contains one or more URLs,} \\ \mathtt{H} \longmapsfrom \text{tweet contains one or more hashtags,} \\ \mathtt{M} \longmapsfrom \text{tweet contains one or more mentions,} \\ \mathtt{D} \longmapsfrom \text{tweet contains one or more medias,} \\ \mathtt{X} \longmapsfrom \text{tweet contains entities of mixed types} \end{array} \right\} \\ &= \{\mathtt{N}, \mathtt{U}, \mathtt{H}, \mathtt{M}, \mathtt{D}, \mathtt{X}\} \end{aligned}$$

Fig. 3.1: Digital DNA representation

A digital DNA sequence is a data representation that is suitable to model the behavior of a single OSN user.However, when analyses are targeted to groups rather than single users, it could be useful to manage and study multiple digital DNA sequences as a whole, in order to infer the characteristics of the group. Here, we study collective behaviors via an analysis of the similarities among the digital DNA sequences of the users of a given group.Among the possible means to quantify similarities between sequential data representations,the longest common substring between two or more DNA sequences. Intuitively, users that share long

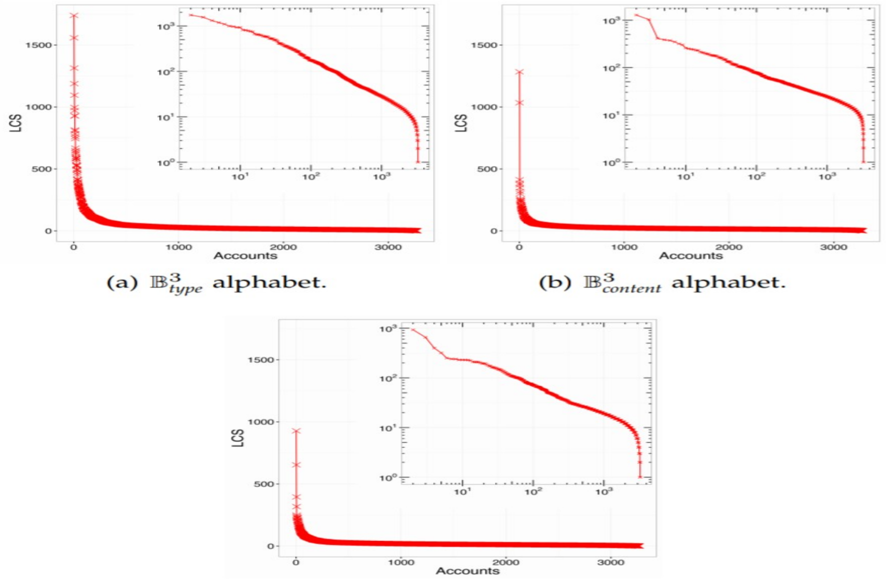
behavioral patterns are much more likely to be similar than those that share little to no behavioral patterns. Given two strings of length n and m , their longest common substring henceforth LCS is the longest string that is a substring of both strings. The extended version of this problem that considers an arbitrary finite number of strings,is called the k-common substring problem.Both the longest common substring and the k-common substring problems can be solved in linear time and space, by resorting to the generalized suffix tree.Fig 3.1 shows the digital DNA representation based on the type and number of bases.

A suffix tree is a compressed tree containing all the suffixes of the given text as their keys and positions in the text as their values. Suffix trees allow particularly fast implementations of many important string operations.Longest common substring problem is to find the longest string that is a substring of two or more strings by using the suffix tree. LCS curves are monotonic nonincreasing functions where K is the length of string with max lenght of Mand is given by

$$LCS[k-1] \geq LCS[k] \quad \forall \quad 3 \ \leq k \ \leq M \qquad \text{(Equ:3.4)}$$

3.3 Characterization of longest common substring curves

To exploit at its best the potential of digital DNA, we need a deeper understanding of the elements that mark the distinction between genuine users and social spambots. Hence, building on the definitions of digital DNA and LCS curves, differences and similarities among those groups of accounts, as seen through the lenses of our digital DNA sequences. In contrast to the remarkably high LCS curves of spambots, genuine accounts show little to no similarity and LCS curves exponentially decay, rapidly reaching the smallest values of LCS length. This preliminary yet considerable differences between the LCS curves of genuine accounts and spambots suggest that, despite the advanced characteristics of these novel spambots,digital DNA is able to uncover traces of their automated and synchronized activity. In turn, the automated behaviors of a large group of accounts results in exceptionally high LCS curves for such accounts. We also consider high behavioral similarity as a proxy for automation and, thus, an exceptionally high level of similarity among a large group of accounts might serve to identify anomalous behaviors. In the following, we preliminarily compare groups of heterogeneous users, looking for features that could be used to design a detection mechanism, while in the next section we detail how to leverage such elements for an effective detection mechanism.The above Fig 3.2 shows the LCS curves of group of accounts.



(a) $\mathbb{B}^3_{type}$ alphabet.

(b) $\mathbb{B}^3_{content}$ alphabet.

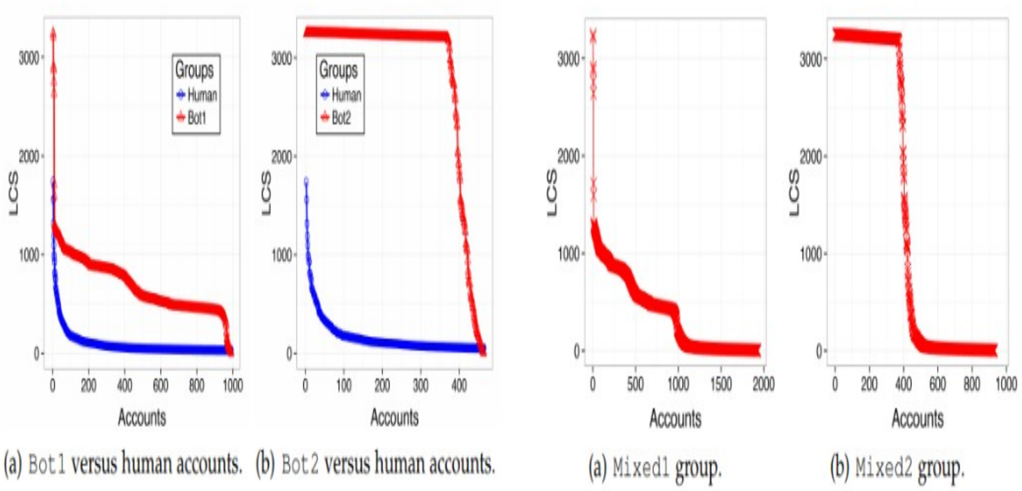
(c) $\mathbb{B}^6_{content}$ alphabet.

Fig. 3.2: LCS curves of a group of genuine accounts

3.3.1 Longest Common Substring curves of a group of heterogeneous users

LCS curves derived from digital DNA sequences of users with similar characteristics, such as genuine Twitter accounts and spambots of a given family. Groups with different characteristics lead to qualitatively different LCS curves. However, we have not yet considered LCS curves obtained from sequences of an unknown and heterogeneous group of users. Thus, leveraging the different groups of accounts studied until now, we built 2 sets of heterogeneous accounts, where we mixed together all the spambots of the Bot1 and Bot2 groups, with an equal number of genuine accounts. Henceforth, such heterogeneous groups of accounts are referred to as Mixed1 and Mixed2, respectively. We observe a continous decrease in the LCS length as the number of considered accounts grows. Such slow decrease is sometimes interleaved by steeper drops, such as those occurring in particlar region. LCS curves in both plots asymptotically reach their minimum value as the number of accounts grows. Overall, such LCS curves show a different behavior than those related to a single group of similar accounts.

Also we have a lack of single trend that spans for the whole domain of the LCS curves. Instead, they depict a situation where a trend seems to be dominant only until reaching a certain threshold. Then, a steep fall occurs and another possibly different trend kicks in. Notably, such portions of the LCS curves separated by the steep drops resemble LCS curves of the single groups of similar users and is used to obtain from a the sets of heterogeneous users, Mixed1 and Mixed2. The steep drops of LCS curves separate areas where the length of the LCS remains practically unchanged, even for significantly different numbers of considered accounts. Plateaux in LCS curves are strictly related to homogeneous groups of highly similar accounts. Note that it is possible to observe multiple plateaux in a single LCS curve. This represents a situation where multiple (sub-)groups exist among the whole set of considered accounts. Furthermore, the steeper and the more pronounced is a drop in a LCS curve, the more different are the two subgroups of accounts split by that drop.Fig 3.3 shows the LCS curves of groups of heterogeneous users.



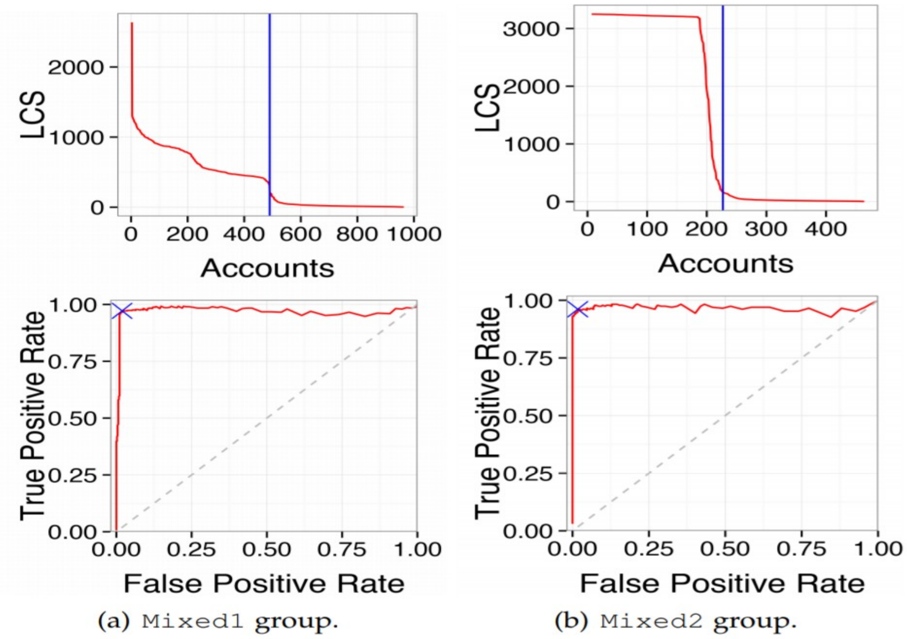
(a) Bot1 versus human accounts. (b) Bot2 versus human accounts. (a) Mixed1 group. (b) Mixed2 group.

Fig. 3.3: LCS curves of group of heterogeneous users

To summarize, LCS curves of an unknown and hetero- geneous group of users can present one or more plateaux, which are related to subgroups of homogeneous users. Conversely, steep drops represent points marking big differences between distinct subgroups. Finally, slow and gradual decreases in LCS curves represent areas of uncertainty, where it might be difficult to make strong hypotheses about the characteristics of the underlying accounts. In conclusion, we argue that LCS curves of an unknown and heterogeneous group of users are capable of conveying information about relevant and homogeneous subgroups of highly similar users.

3.4 Social Fingerprinting

Social Fingerprinting refers to the ways detection of the two subgroups of spambots and genuine accounts that constitute our Mixed1 and Mixed2 groups.Two different methods to split the ac- counts of the Mixed1 and Mixed2 groups, according to the characteristics of their LCS curves are done. Approaches based on supervised and unsupervised approach, showing the suitability of LCS curves and also the effectiveness of the detection mech- anisms. With both the approaches, we consider as spambots those accounts that are related to high LCS values, namely sharing long behavioral patterns. Consider as genuine users those accounts that share little portions of their digital DNA.provide a rigorous assessment of the possibility to detect spambots using the LCS curves of groups of heterogeneous accounts.



(a) Mixed1 group. (b) Mixed2 group.

Fig. 3.4: Supervised learning LCS , ROC curves

3.4.1 Finding subgroups of similar users: a supervised approach

In the spambot detection supervised approaches are commonly employed to discriminate between spambots and genuine users. Supervised classifiers start analyzing a training-set, where the class of every user is specified in order to understand the characteristics of the two classes of users. Then, they exploit such learned characteristics to automatically discriminate between spambots and genuine users in a new set of unlabeled users. In addition, a test-set is used to evaluate and compare the effectiveness of different classifiers. This approach is typically performed in all kinds of machine learning classification tasks.

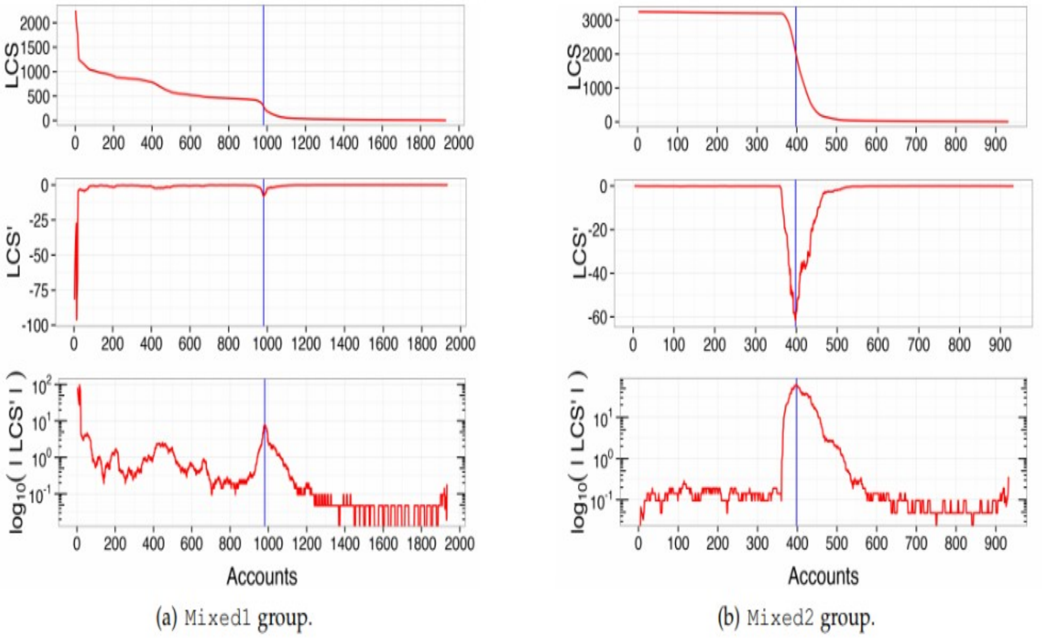
A methodology to combine LCS curves and user labels available from a training-set was used as a supervised approach for the detection of subgroups of users. A good division of the original set of users into several subgroups is one where all the users belonging to a given class are assigned to the same subgroup.In theory, any point of the LCS curve of a heterogeneous group of users can be used as a splitting point to obtain two sub- groups of more homogeneous users. However,it is not all possible to find splitting points which lead to accurate subgroup partitioning. Using the given labels, we can evaluate every possible splitting point in the LCS curve of the training- set users and find the one that yields the best possible subgroup division. To this regard, every point generates a different classifier that can be evaluated in terms of machine learning performance metrics. The LCS value associated to the classifier that achieves the best results, according to a given metric of choice, is then used as a threshold to classify users of the test-set. The different classifiers can also be qualitatively evaluated by means of ROC curves, where the best classifiers are those that lay near to the top-left corner of the plot. The diagonal line in ROC curves is instead related to a random classifier[4].

Fig 3.4 shows the LCS curve splitting point and ROC curves of classifier. Among the various metrics commonly adopted to evaluate machine learning classifiers, we picked the best classifier as the one achieving the highest Matthews Correlation Coefficient (MCC). In ROC curves those users laying to the left of the vertical splitting line as spambots, and those other users laying to the right of the vertical splitting line as genuine.

Notably, as usual for classification approaches that operate in a supervised fashion, one cannot guarantee that the learned LCS value would still be effective when applied on a test-set different from the one used to derive such LCS value. This problem is known in machine learning literature as transfer learning or inductive learning [5]. In order to overcome this limitation, in the following section we define an unsupervised approach for discriminating between spambots and genuine users, that does not suffer from this drawback.

3.4.2 Finding subgroups of similar users: an unsupervised approach

An unsupervised methodology that leverages previous findings and exploits the shape of LCS curves of heterogeneous users in order to find subgroups of users with similar behaviors. Specifically, we propose to exploit the discrete derivative of a LCS curve to recognize the points corresponding to the steep drops. This approach is applicable to a broad range of situations, since it requires no information other than the LCS curve of the heterogenous group of users.



(a) Mixed1 group.

(b) Mixed2 group.

Fig. 3.5: Unsupervised learning LCS curves

The steep drops of LCS curves appear as sharp peaks in the derivative plot and represent suitable splitting points to isolate different subgroups among the whole set of users. All the suitable splitting points might be ranked according to their corresponding derivative value and then, a hierarchical top-down approach may be applied, by repeatedly dividing the whole set of users based on the ranked points, leading to a dendrogram structure. For instance, this approach can be exploited in situations where the LCS curve exhibits multiple plateaux and steep drops, in order to find the best possible clusters that can be used to divide the original set of heterogeneous users.

The discrete derivative of the LCS curve , given that LCS curves are monotonic nonincreasing functions and their derivatives will assume only zero or negative values, with steep drops in the LCS corresponding to sharp neg- ative peaks in derivative if LCS . Simple peak-detection[6] algorithms can be employed in order to automatically detect the relevant peaks in derivative of LCS graph. Notably, this approach does not require a training phase and can be employed pretty much like a clustering algorithm, in an unsupervised fashion. To prove the effectiveness of this unsupervised approach, we applied it to the LCS obtained from the unlabeled Mixed1 and Mixed2 groups, with the goal of sepa- rating spambots from genuine users. The logarithmic scale plots of the derivatives have been computed and they have been added for the sake of clarity, since they highlight the less visible peaks of the linear scale plots. In order to facilitate the detection of peaks in derivative graph , we smoothed the original LCS curves before computing their derivatives. This preprocessing step acts pretty much like a low-pass filter, allowing to flatten the majority of noisy fluctuations.

The proposed methodology accurately identified reasonable splitting points in order to find two clusters among the whole sets of unlabeled users. In detail, those users laying on the left of the vertical splitting line that is, users sharing long behavioral patterns long LCS are labeled as spambots. Conversely, the users to the right of the vertical splitting line, users sharing little similarities are labeled as genuine ones. Together with this qualitative assessment, quantitative evaluation of our spambots detection techniques, by means of well-known performance metrics of machine learning algorithms. We remark that, although the Mixed1 and Mixed2 groups feature an equal number of genuine and spambot accounts, the ratio between the two types of accounts is generally different when considering the whole Twitter. The balance in Mixed1 and Mixed2 is because we mostly envisage the application of our digital DNA.Fig 3.5

shows the lCS curves used in unsupervised methods.

DNA technique to spot anomalous groups within devoted events,campaigns are those accounts retweeting a specific hashtag, or participating in an electoral campaign, or which are followers of a certain account. Whereas the analysis is concentrated on a subset of accounts acting around a par- ticular event, the ratio between genuine and spam accounts can drastically vary, even leading to a balance in the cardi- nality of the two groups.

3.4.3 Comparison of the two approaches

In Both the supervised and the unsupervised approaches identified similar splitting thresholds, that lay inside the steepest drops of the LCS curves of Mixed1 and Mixed2. However, results of the supervised and unsupervised approaches are slightly different, especially with regards to the accounts of the Mixed2 group. In the following, we provide a quantitative comparison of the two approaches to assess which one actually better discriminated between spambots and genuine users.

To summarize the outcomes of the supervised and the unsupervised approaches, we leverage evaluation metrics based on four standard indicators True Postive,True Negative,False Postive and False Negative which forms the confusion matrix.

We obtain the performance measure based on the above confusion matrix values and they are Accuracy,Precision,Recall,Specificity,F-measure,Matthews Co-relation Coefficient.Accuracy measures how many users are correctly classified in both of the classes, but it does not express whether the positive class is better recognized than the other one. Moreover, there are situations where some predictive models perform better than others, even having a lower accuracy. A high Precision indicates that many of the users identified as spambots are indeed real spambots, but it does not give any information about the number of spambots that have not been identified as such. This information is instead provided by the Recall metric, indeed a low Recall means that many spambots are left undetected. Specificity instead measures the ability in identifying genuine users as such. Finally, F-Measure and MCC convey in one single value the overall quality of the prediction, combining the other metrics.

Finally, we evaluated the performances of the unsuper- vised approach with a dataset that reflects a real word scenario, where the number of spambots is supposed to be much smaller than the number of human-operated ac- counts. Then, we randomly picked the DNA sequences of the two original test-sets so as to build mixed datasets with the correct numbers of spambots and genuine accounts. Finally, we executed the unsupervised detection approach on such datasets and evaluated the detection performance, averaging the results over 20 executions. From the plot it is noticeable that the performance improves as the number of bots in the dataset increases. Considering that the number of spambot accounts in this experiment is extremely low the reliability of the unsupervised approach is still noticable.

The Matthews Correlation Coefficient (MCC) has a range of -1 to 1 where -1 indicates a completely wrong binary classifier while 1 indicates a completely correct binary classifier. Using the MCC[7] allows one to gauge how well their classification model/function is performing.Being a correlation coefficient, MCC 1 means that the prediction is very accurate, MCC 0 means that the prediction is no better than random guessing, and MCC 1 means that the prediction is heavily in disagreement with the real class.Nonetheless, also the unsupervised approach is able to provide overall accurate predictions.

3.5 Performance evaluvation and comparison

Social Fingerprinting technique, we compared our detection results with those obtained by different state-of-the-art spambot detection techniques, namely the supervised one by Yang et al[8], and the unsupervised approaches by Miller et al[9] and by Ahmed et al. The work presented provides a machine learning classifier that infers whether a Twitter account is genuine or spambot by relying on accounts relationships, tweeting timing, and level of automation. Instead, works in Miller et al and by Ahmed et al[10] define a set of machine learning features and apply clustering al- gorithms. Specifically, in Miller et al the authors propose modified versions of the DenStream and StreamKM++ algorithms respectively based on DBSCAN and k-means and apply them for the detection of spambots over the Twitter stream. Ahmed et al. Ahmed et al exploit the Euclidean distance between feature vectors to build a similarity graph of the accounts and graph clustering and community detection algorithms to identify groups of similar accounts in the graph.

3.5.1 Existing evaluation techniques

Density-based spatial clustering of applications with noise DBSCAN[9] is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jrg Sander and Xiaowei Xu in 1996.It is a density-based clustering algorithm given a set of points in some space, it groups together points that are closely packed together points with many nearby neighbors, marking as outliers points that lie alone in low-density regions whose nearest neighbors are too far away. DBSCAN is one of the most common clustering algorithms.

StreamKM++ [10]computes a small weighted sample of the data stream,called the coreset of the data stream. A new data structure called coreset tree is developed in order to significantly speed up the time necessary for sampling non-uniformly during the coreset construction. After the coreset is extracted from the data stream, a weighted k-means algorithm is applied on the coreset to get the final clusters for the original stream data. First, we use an adaptive, nonuniform sampling approach similar to the kmeans++ seeding procedure to obtain small coresets from the data stream. This construction is rather easy to implement and, unlike other coreset constructions, its running time has only a small dependency on the dimensionality of the data. Second, we propose a new data structure, which we call coreset tree. The use of these coreset trees significantly speeds up the time necessary for the adaptive, nonuniform sampling during our coreset construction.

3.5.2 Emerging novel spambots

Social fingerprinting work tackled the detection of a novel wave of social Twitter spambots. By accurately mimicking the characteristics of genuine users, these spam- bots are intrinsically harder to detect than those studied by Academia in the past years. As a consequence, it is exceptionally difficult to detect such spambots working on an account by account basis, as in the case of machine learning classifiers. This claim is supported by the poor detection results obtained by the approach of Yang et al. Spambots did not evolve in the way that Yang et al. imagined. In turn, our work also provides additional evidence of the emergence of a new wave of spambots, Despite the advanced characteristics of these new spam- bots, we argue that the traces of their automated nature are still present in the history of their behaviors. Such subtle traces might not be enough to infer the nature of an account whether genuine or spambot by simply analyzing his past behaviors.

Nonetheless, they can be leveraged by observing collective behaviors of groups of accounts. Since spambots of same family that is, those spambots belonging to the same botmaster and perpetrating the same illicit activity must necessarily have the same goal, and hence similar behaviors, it is possible to exploit behavioral similarities between large groups of accounts as a proxy for automation. Indeed, the proposed technique features the ability to uncover those characteristics that are typical of a group of similar or synchronized accounts.

3.5.3 Complexity and scalability

Notably, the best performing techniques proposed in recent years for spambots detection are based on data and time demanding analyses. This highlights a trade-off between accuracy and responsiveness of spambots detection. The amount of data needed to calculate features and the resulting lack of responsiveness in providing results also un- dermine the large-scale applicability of such detection techniques. To this regard, our Social Fingerprinting technique is not only effective, but also efficient. Indeed, some of the previously mentioned approaches for spambots detection are among those requiring a large number of data-demanding features and computationally demanding algorithms.

For instance, approaches that are based on graph mining have been proved to be more demanding in terms of data that is needed in order to perform the detection. In addition, to address and solve other research challenges in the field of social networks, other algorithms for the analysis of biological DNA and strings can be drawn from the established literature in the fields of bioinformatics and string mining. . In particular, we monitored the effects of increasing the number of investigated accounts, increasing the length of their digital DNA sequences, and changing the considered digital DNA alphabet , on execution time and memory consumption. Considering that the LCS problem is still well studied and has several solutions leveraging high parallelization in distributed computing environments, we can conclude that our approach is scalable enough and that it can be adopted to deal with real cases.

3.5.4 Flexibility and multidimensionality of digital Deoxyribonucleic acid

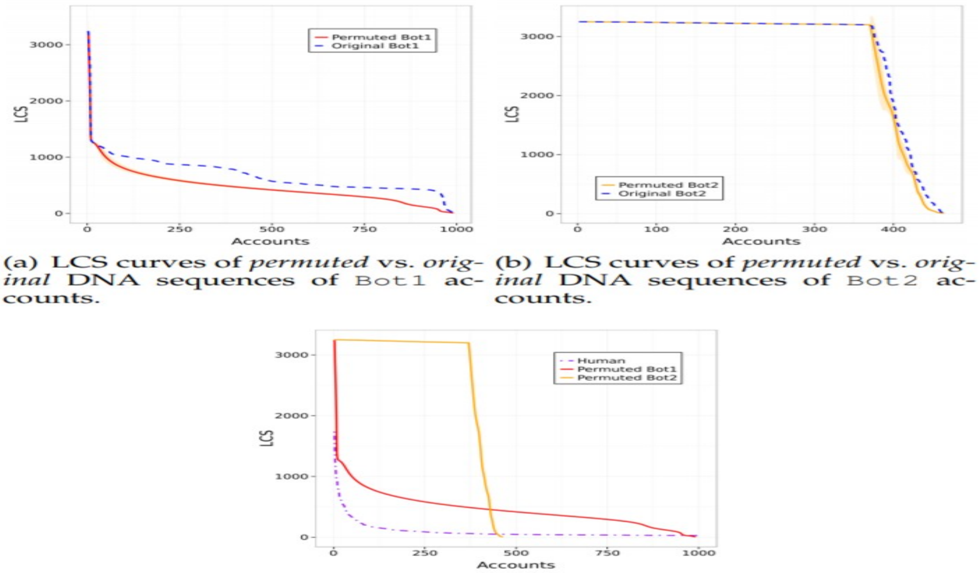
Provided the great level of flexibility of digital DNA, we also envision the possibility to exploit results of our Social Fingerprinting technique as a feature in a more complex detection system. For instance, a hybrid detection system leveraging features derived from the digital DNA analysis and other machine learning features, or a system that simultaneously exploits multiple types of digital DNA. Then, results of these models could be used simultaneously in an ensemble or voting system. However, the same accounts might instead be unambiguously characterized by the LCS curve obtained with a different digital DNA alphabet.

Hence, exploiting multiple alphabets might allow to uncover more characteristics of the accounts under investigation, ultimately leading to better detection results. Further alphabets according to which it could be possible to extract the digital DNA of online users are as follows. One alphabet could capture the interaction patterns of Twitter users, considering the popularity level of the peers with whom a given user interacts. Specifically, we may think to exploit retweets and replies among users as a form of interaction and an accounts followers count as a measure of popularity for that account.

Finally, we notice how our proposed approach is very generic and flexible, since it makes possible to deepen the analysis of LCS curves with the straightforward use of powerful tools, such as dendrograms for clustering and ROC curves for classifiers, which are already and widely adopted in a number of machine learning tasks.

3.5.5 Defense against evading techniques

Given the focus of the Social Fingerprinting technique on the sequence of user actions, one could foresee evading spam- mers to randomly re-order the sequence of their tweets, in an effort to escape detection. In order to thoroughly assess the impact of such evading technique on our detection performances, we have run a series of experiments via Monte Carlo simulations. On the one hand, spambots have less variability in their sequences and they tend to have a DNA base that is predominant with respect to the other ones. On the other hand, genuine accounts feature a base distribution that is almost uniform. This explains why, in our data, randomly reordering the sequences of spambots, although partly erasing similar behavioral patterns, still allows to distinguish them from humans, according to our LCS similarity approach.



(a) LCS curves of *permuted* vs. *original* DNA sequences of `Bot1` accounts.

(b) LCS curves of *permuted* vs. *original* DNA sequences of `Bot2` accounts.

(c) LCS curve of genuine (human) accounts and those of permuted `Bot1` and `Bot2`.

Fig. 3.6: Monte carlo approach

Monte Carlo situation generally refers to the test data obtained by permuting the base strings of users. Monte Carlo simulation proved to be surprisingly effective at finding solutions to these problems. Since Monte Carlo methods have been applied to an incredibly diverse range of problems in science, engineering, and finance and business applications in virtually every industry.Monte Carlo methods can be used to solve any problem having a probabilistic interpretation. By the law of large numbers, integrals described by the expected value of some random variable can be approximated by taking the empirical mean of independent samples of the variable. When the probability distribution of the variable is parametrized, mathematicians often use a Markov chain Monte Carlo sampler. The central idea is to design a judicious Markov chain model with a prescribed stationary probability distribution. That is, in the limit, the samples being generated by the method will be samples from the desired distribution. By the ergodic theorem, the stationary distribution is approximated by the empirical measures of the random states of the sampler.Fig 3.6 shows the LCS curves of the permuted account base string to the genuine ones.

Humans hired to perform spamming tasks. While crowdsourcing spammers may demonstrate a certain level of similarity, they probably manifest behaviors that are not as consistent as the ones of automated accounts. As such, detection mecha- nisms targeting crowdsourcing spammers may benefit from leveraging the more flexible longest common subsequence.

Conclusion

Social spambots in Open Social Networks whose recent waves of spam- bots have been thoroughly engineered so as to mimic the human behavior of OSNs genuine users. These novel species of spambots do escape state-of-the-art algorithms specifically designed to detect them.The proposed digital DNA behavioral modeling technique leveraging this methodology, it was able to distinguish the low intensity signals that make humans different from bots, when considering users not on an account by account basis, but rather on collective behaviors. Social Fingerprinting detection approach and coupled algorithmic toolbox drawn from the bioinformatics and string mining domains have shown excellent detection capabilities for all of the most relevant detection metrics, outperforming state-of-the-art solutions.

References

[1] G. Stringhini, C. Kruegel, and G. Vigna, Detecting spammers on social networks, in 26th Annual Computer Security Applications Conference (ACSAC). ACM, 2010, pp. 19.

[2] K. L. Gwet, Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC, 2014.

[3] R. Zafarani, M. A. Abbasi, and H. Liu, Social media mining: an introduction. Cambridge University Press, 2014.

[4] T. Fawcett, An introduction to ROC analysis, Pattern recognition letters, vol. 27, no. 8, pp. 861874, 2006.

[5] S. J. Pan and Q. Yang, A survey on transfer learning, IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 13451359, 2010.

[6] V. Lampos and N. Cristianini, Nowcasting events from the social web with statistical learning, ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3, no. 4, p. 72, 2012.

[7] P. Baldi, S. Brunak, Y. Chauvin, and H. Nielsen, Assessing the accuracy of prediction algorithms for classification: An overview, Bioinformatics, vol. 16, no. 5, pp. 412424, 2000.

[8] 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, pp. 12801293, 2013.

[9] F. Ahmed and M. Abulaish, A generic statistical approach for spam detection in online social networks, Computer Communications, vol. 36, no. 10, pp. 11201129, 2013.

[10] Z. Miller, B. Dickinson,W. Deitrick,W. Hu, and A. H.Wang, Twitter spammer detection using data stream clustering, Information Sciences, vol. 260, pp. 6473, 2014.