Nationality and Geolocation-Based Profiling in the Dark(Web)

Massimo La Morgia, Alessandro Mei, Eugenio Nerio Nemmi, Simone Raponi, and Julinda Stefa

Abstract—In this paper we are concerned with geolocating the anonymous crowds of Dark Web forums. We do not focus on single users, but on the crowd as a whole. We work in two directions: The first idea is to exploit the time of all posts in the Dark Web forums to build profiles of the visiting crowds and to match the crowd profiles to that of users from known regions. Then, we develop a new dataset to detect the native language of the crowds to support and integrate this match. We assess the effectiveness of our methodology on the standard web and two Dark Web forums with users of known origin, and apply it to three controversial anonymous Dark Web forums. We believe that this work helps the community better understand the Dark Web from a sociological point of view and supports the investigation of authorities when the security of citizens is at stake.

Index Terms—Dark web, distributed systems, anonymity, native language identification.

1 Introduction

The Dark Web hit the news eight years ago with the rise of Silk Road. On Silk Road—a clandestine drug market hidden in the Dark Web—Internet users of Tor [1] could use bitcoins to buy all sorts of psychedelics with excellent anonymity. A couple of years later, in 2013, the founder of Silk Road was arrested and the site taken down. At that point it was an estimated $1.2 billion business, and, after it was shut down, countless successors quickly proliferated.

In the popular culture, the Dark Web is associated with criminal activities—drug sale, identity theft, money laundering, computer hacking, botnets, credit card frauds, gun sales, child pornography, and other related cyber-crimes. This is only partly true, anonymity technology like Tor and Bitcoin were designed as a product of debates among technology libertarians in the past decades and Tor and the Dark Web are actually important to support freedom of information and speech in the Internet, especially in countries where the government or other powerful entities try to suppress it. Indeed, an important part of the Dark Web is made of forums where people can debate any matter of interest. Often, these forums are about topics that are illegal, controversial, or considered questionable by the society. In other cases, they are meeting places where dissidents of authoritarian countries can freely discuss politics without being censored or prosecuted. Examples of political sites in the Dark Web are Strongbox or GlobaLeaks. Strongbox is promoted by the Freedom of the Press Foundation, while GlobaLeaks by The Hermes Center for Transparency and Digital Human Rights. Examples of forums about questionable topics are the CRD Club, a Russian site on computer hacking and technology frauds, or the Dream Market, a forum about the quality of drugs and vendors in the associated marketplace.

In this paper, we consider the problem of uncovering the geographical origin of the crowds of Dark Web forums. Forums are one of the most important Dark Web services, and privacy is essential. We do not attack the anonymity of the single forum visitor, we are interested in understanding the geographical distribution of the visitors as a collective property. In these respect, this paper considers the notion of anonymity in the Dark Web with a new angle. Previous work, especially on Tor, has focused on attacking anonymity mostly by using traffic analysis or web browser fingerprinting. In the first case, the adversary controls both the endpoints in the Tor mixing circuit, or even the autonomous systems of the entry and exit points of the circuit, and is able to de-anonymize a single user by correlating the traffic at the two endpoints. In the second case, the adversary controls the local network of the user and is able to understand the destination site of her browsing session by fingerprinting the traffic and matching the fingerprint against a set of known web sites. In our work, we do not assume any control of the network, we just access the forum and analyze the messages and the profile of access to the forum as documented by the site—information that is available to every member of the forum with no particular privilege.

We introduce two methodologies. In the first, we show how to decompose the global profile of posting of the Dark Web forum into components that uncover the geographical origin of the crowd. The fundamental idea is to consider the time of all postings and match it to the profile of Internet activity on standard web forums of crowds from known regions. The second methodology focuses on detecting the native language of anonymous Dark Web users, starting from their posts in English. We build two classifiers: one to detect the English native speakers and the second one to identify the native language of the non-native English speaker, reaching an F1-score of 82.2% and 81.6% respectively. To achieve this second goal, we build a new dataset, specific for web slang. Finally, we combine these two methodologies to obtain fine-grain profiling of the crowds of Dark Web forums.

We validated the first methodology with experiments in

- *M. La Morgia, A. Mei, E. N. Nemmi, J. Stefa are with the Department of Computer Science, Sapienza University of Rome, Italy.*
  Email: {lamorgia, nemmi, mei, stefa}@di.uniroma1.it.
- *S. Raponi is with Hamad Bin Khalifa University (HBKU), CSE-ICT Division, Doha, Qatar. Email: sraponi@mail.hbku.edu.qa.*

real forums in the Dark Web. The first two, the CRD Club, which is in Russian, and the Italian DarkNet Community (IDC), confirmed our findings. They are known to be centered in Russia and Italy, respectively, as correctly predicted by our methodology. Then, we uncovered the crowds of the Dream Market, The Majestic Garden, a site of fans of psychedelic experiences, and the Pedo Support Community, a forum on child abuse. According to our analysis, the first site is mostly European (with an important component from North America); the second one is mostly North American (with a smaller component from Europe). The third one, arguably the most controversial, has an essential component of the crowd living in Southern Brazil or Paraguay. Then, we added more information about the crowds with our methodology of native language identification.

We believe that our contribution can be key to better understand the Dark Web and its plethora of forums from a sociological point of view. Not only that, our methodology can give important initial information on the geographical origin of the users of a particular forum and, in case of illicit activities, support the discovery of their real identities by using known de-anonymization techniques in the autonomous systems of the regions where most of them live.

2 Background

Tor [1] is one of the most popular anonymity systems. With over 2 million users, about $7,000$ relays, $3,000$ bridges, and $50,000$ estimated hidden services, it is also one of the largest. Tor can be used to access the Internet anonymously and to use services that are unreachable due to, for example, censorship. The main idea is that the user selects a circuit that typically consists of three relays—an entry, a middle, and an exit node. The user negotiates session keys with all the relays and each packet is encrypted multiple times, first with the key shared with the exit node, then with the key shared with the middle node, and lastly with the key shared with the entry node (also known as the guard). To send a packet to the final destination anonymously, it is first sent to the guard. The guard removes the outer encryption layer and it relays the packet to the middle node. In turn, the middle node removes its encryption layer and relays the packet to the exit node. Lastly, the exit node removes the last layer of encryption and relays the packet to its final destination. Thanks to Tor, the user can get anonymous access to Internet services like standard websites, for example.

Tor is also known in the Internet community as one of the core infrastructure to access the Dark Web. The Dark Web is the set of online web resources that are not indexed by common search engines and that can not be explored without using anonymity technologies such as Tor, I2P [2], or Freenet [3]. Technically, the services that run in the Dark Web under Tor technology are called hidden services. Hidden services have their own top level domain which is .onion, and their host name consists of a string of 16 characters derived from the service's public key. To keep mutual anonimity, both the user and the hidden service (the website) set up independent Tor communications to a common rendez-vous point, chosen with the help of specific directory services. This way, both entities are anonymous to each other and to every other node in the network.

3 Time-Zone Geolocation of Crowds

Our behavioral patterns, including access to websites or Dark Web hidden services, is affected by our everyday life rhythm. During the day we engage in activities in a systematic way mostly dictated by the local time and daylight—waking up, going to work or school, having lunch, possibly doing afternoon activities, having dinner, resting. This is confirmed in [4], [5], where the authors analyzed Facebook and YouTube access patterns. In both services, the requests steadily grow from the early morning to the afternoon with a peak between 17:00 (5pm) and 22:00 (10pm), then the number of requests drops rapidly during the night. In this line, our idea is to use the correlation between the everyday life rhythm (timezone and daylight) and the access or post patterns of users of forums in the Dark Web to uncover their location in terms of timezone. The first step is to generate access profiles that are common to users of a certain geographical region (e.g. nation). We do so for several regions of the planet. Then, given the access profile of a crowd of users of which we know nothing of, we uncover their origin according to the similarity with known profiles.

3.1 Building Reliable User and Region Profiles from User Activity Traces

In this section we show how we build profiles of users from a given known population starting from their activity traces. The traces can be of any kind: posts, comments to posts, messages exchanged, access times, or a mix of them. We focus on building profiles that describe the level of online posting activity of the population throughout the day. We start by profiling single users. In particular, we determine whether a user is or is not typically active at a given hour of the day. For this reason, the profile P_u of user u is represented by an array of 24 elements, one per hour—element $P_u[h]$, $h \in \{0, \ldots, 23\}$, is the fraction of daily online posting activity done by user u during hour h. Let boolean $a_u(d,h)$, indicate whether user u has posted in the h^{th} hour of day d. The profile P_u is then defined as follows:

$$P_u = \{P_u[h] | h \in \{0, \ldots, 23\}, P_u[h] = \frac{\sum_d a_u(d,h)}{\sum_{d,h'} a_u(d,h')}\}. \tag{1}$$

Intuitively, profile P_u is the distribution of user u activity throughout the day on the target forum. The overall population profile P is an aggregate of all user profiles as follows:

$$P = \{P[h] | h \in \{0, \ldots, 23\}, P[h] = \frac{\sum_u P_u[h]}{\sum_{u,h'} P_u[h']}\} \tag{2}$$

To build reliable region profiles we need to start off from datasets that are rich enough to reflect the behavioral patterns of the users and that include verified information on their location. One possibility is the dataset [6] obtained from the Twitter livestream representing around 2% of the total Twitter streams in 2016. Includes tweets of $6,058,635$ users all over the world whose home country is retrievable from their Twitter profile. Using this dataset and the above methodology we have built profiles for 14 countries or states: Brazil, California, Finland, France, Germany, Illinois, Italy, Japan, Malaysia, New South Wales (Australia), New

TABLE 1
Twitter dataset—active users by Country/State.

Country/State	Users (#)	Country/State	Users (#)
Brazil	3,763	Japan	3,745
California	2,868	Malaysia	1,714
Finland	73	New South Wales	151
France	2,222	New York	1417
Germany	470	Poland	375
Illinois	794	Turkey	1,019
Italy	734	United Kingdom	3,231

York, Poland, Turkey, and the UK. To do so, we have considered daylight saving time for all corresponding regions and we have filtered out periods of particularly low activity, like holidays. In addition, we have also filtered out non active users—users with just a handful of posts, lower than a certain threshold, that do not give enough information to profile their behavior in the long run.We chose the threshold of 30 posts, as we noticed that it is a reasonable value to get a meaningful profile. Table 1 shows the regions considered along with the number of active users.

As an example, we show in Figures 1(a) and 1(b) the profiles of a random German user and of the German population, respectively. First, we note that in both profiles we can easily distinguish the night as the hours of lower activity (the interval between 1:00 (1am) and 7:00 (7am)). In addition, we can observe that activity of the German user in Figure 1(a) has a first peak in the morning, drops during lunch time, and starts to grow again from the early afternoon to the evening, following a typical daily rhythm.

It is important to note that the profile of the German population follows the same pattern that has been found in Facebook and YouTube [4], [5]. Actually, this is true for all the populations of the countries we have considered in Table 1. In other words, when the profile of large crowds coming from different timezones are brought to the local time, their profiles are almost identical. To confirm this observation we have computed the Pearson correlation for every pair of countries or states in Table 1, and the value is constantly higher than 0.9. Therefore, we can use a generic profile independently of the region or nationality, after the user activity is properly shifted to the local time. As an example, we have plotted the profile of the entire Twitter dataset in Figure 1(c). Note how the profile is very close to the one of the German population, with the only difference of 1 hour shift. For example, the evening peak of activity is at 22:00 (10pm) UTC$+1$) for the German crowd, just like the general Twitter profile that has the peak at 21:00 (pm) UTC. Therefore, we can consistently use the general profile as the common baseline, properly shifted to the right timezone.

3.2 Placing Anonymous Users to Time Zones

Users of the same region typically have a profile that is very close to that of the corresponding timezone crowd, and further away from crowds of different timezones. So, for every member of an anonymous crowd, we compare his profile with that of all different timezone profiles built with the method described in the previous section. Then, we geolocate that member to the timezone whose activity profile is less distant: The one for which it takes less effort to transform the single user profile into by both shifting and moving probability mass. (Recall that activity profiles are activity distributions). An adequate distance measure in this view is the Wasserstein metric [7], also known as the Earth Mover's Distance (EMD). Given two distributions of earth mass spread on the same space, the EMD measures the least amount of work to move earth around so that the first distribution matches the second.

3.3 Single-Country Placement

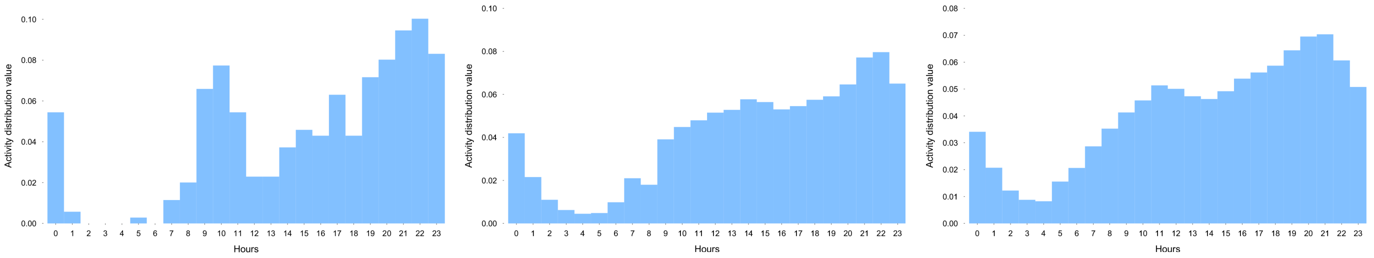
To assess the accuracy of our geolocation methodology, we first apply it to the Twitter dataset, enriched with ground truth information. We start off with the Germany. For every timezone, we compute the fraction of the population with profiles falling into Germany's timezone according to the EMD. Despite common nationality, the habits of two different people are not exactly the same. For example, youngsters tend to go to sleep later than older people, parents wake up earlier than teenagers, and so on. This should also be reflected in their activity profiles. So, while we expect a large number of the German crowd to fall under the timezone of Germany, we also foresee that a portion of the crowd will be placed in neighbor timezones. This is confirmed by Figure 2(a), which plots the percentage of Germans placed to the 24 timezones according to the EMD. We first observe that there is a peak at UTC $+ 1$ timezone, that covers Germany, while the values drop for timezones further away. Most importantly, we observe that the crowd placement follows a Gaussian distribution, with a standard deviation between the fitted Gaussian and the crowd distribution of 0.013.

Figures 2(b) and 2(c) show the distributions for the populations of France and Malaysia, respectively. Again, we observe that they follow a Gaussian distributions centered in the timezone of the corresponding country. The same trend holds for all the other countries in Table 1, whose graphs we omit due to space limitations. It is worth mentioning that, after applying curve fitting [8] to the distributions, we note that the x axis value corresponding to the peak of the placement matches the mean of the Gaussian distribution. We also found that the average Gaussian standard deviation value for all the countries considered is $\sigma \simeq 2.5$.

These observations bring us to the conclusion that, to geolocate a given crowd of people from the same, unknown region, it is enough to build the corresponding activity profiles placement through the EMD distance and curve-fit the resulting distribution with a Gaussian. The center of the Gaussian will uncover the timezone of the unknown region and thus the geolocation of the crowd.

3.4 Multiple-Country Placement

Oftentimes, users access a given site from multiple different regions. Since single region crowds follow a Gaussian distribution, we expect that the mixture of multiple region populations exhibits a profile that follows a Gaussian mixture model. Thus, uncovering the Gaussian distributions (i.e. mean and standard deviation) allows us to correctly place the members of mixed-country crowds in the corresponding geolocations. However, this is not an easy task. The reason is that we do not know a priori the number of different

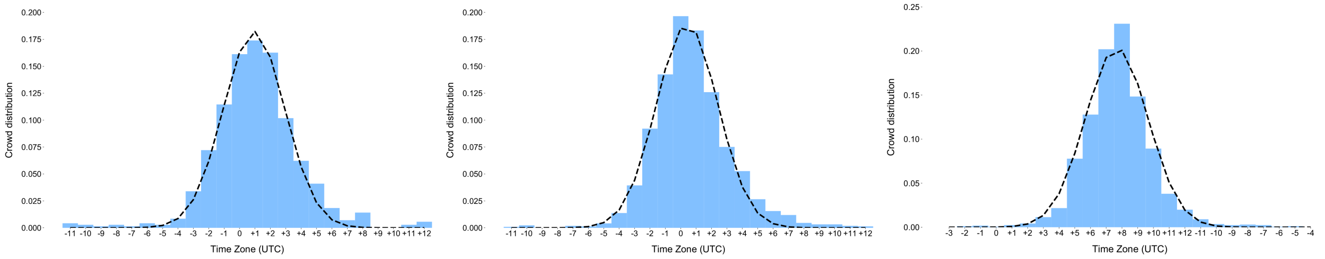


(a) Profile of a German user.

(b) Twitter dataset of the German population (local time UTC + 1).

(c) Entire Twitter dataset (UTC).

Fig. 1. Profiles on the Twitter Dataset: Single German profile vs German (UTC + 1) vs Generic profile (UTC).



(a) German

(b) French

(c) Malaysian

Fig. 2. EMD based placement of national crowds.

regions of the crowd. To address this issue, we initialize the Expectation-Maximization [8] (EM) with the standard deviation $\sigma \simeq 2.5$ observed empirically for the Gaussian fitting curves of single-region placement distributions in the previous section. EM is an iterative algorithm used as the standard to estimate the maximum likelihood parameters of a given model. In our case, the model is the Gaussian mixture, and the components are the Gaussian curves. Dempster et al. [9] show the effectiveness of the EM to estimates the parameters for finite mixtures of parametric families.

We test the effectiveness of the Gaussian Mixture Model (GMM) based geolocation with the Twitter dataset, on which we have ground-truth information regarding the nationality of the users. This time we build two synthetic distributions of multiple-region crowds as follows. The first synthetic distribution is made of a three-way repetition of the Malaysian user activity according to three different timezones: UTC, Californian (UTC − 7), and the Australian region of New South Wales (UTC + 9). In the second distribution we simply merge together users from different regions: Illinois (UTC−6), Germany (UTC+1), and Malaysia (UTC + 8). The results of the geographical classification for both cases are shown in Figures 3(a) and 3(b). Note that we accurately uncover both the number of regions per crowd given by the number of Gaussian curves and the corresponding timezones that match the Gaussian distributions.

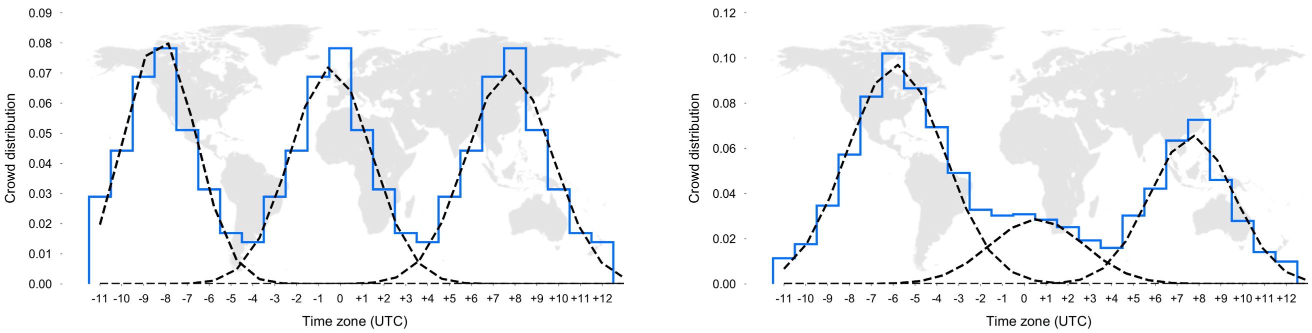
Lastly, in order to quantify how well the fitted Gaussians match the crowd distributions, we have computed the average and standard deviation of the point-by-point distance of the two (see Table 2 for all graphs included in this paper). As benchmark we computed the same metrics for the Malaysian dataset with the corresponding Gaussian fitting shifted of 12 hours (last row of the table). We note that both metrics are very low for both the single-country fitting (first three rows) and the multiple-country fitting (fourth and fifth row of the table). This is particularly true when we compare them to the baseline values, suggesting that the Gaussian curves fit well the crowd distribution.

TABLE 2
Gaussian fitting metrics.

Dataset	Average	Standard deviation
Malaysian Twitter	0.009	0.013
German Twitter	0.009	0.009
French Twitter	0.008	0.010
Synthetic dataset (a)	0.011	0.010
Synthetic dataset (b)	0.012	0.010
CRD Club	0.007	0.006
Italian DarkNet Community	0.014	0.016
Dream Market forum	0.011	0.008
The Majestic Garden	0.009	0.011
Pedo support community	0.012	0.010
Baseline	0.081	0.070

3.5 Polishing the Datasets

The EMD is also used to filter out users with so called flat profiles: Users whose activity profile are very close to being uniformly distributed over all the hours. From an in-depth investigation on the Twitter dataset we saw that these kind of users are typically bots. At any rate, the flatness of their profile makes so that there is no meaningful information that distinguishes them from a bot machine. In addition, they do not contribute in a meaningful way to the creation of timezone profiles. Thus, we have decided to remove these profiles from the datasets. To do so, we remove all the users whose profiles, according to the EMD, result being closer



(a) Synthetic dataset modelling the behavior of Malaysian users in three different timezones: UTC, California, and Australia.

(b) Synthetic dataset: Illinois, German, and Malaysian users.

Fig. 3. Geographical classification of multiple-region crowds.

to an artificial profile created by us where every value is of $1/24$ $(1/(\#\text{daily hours}))$ than to a timezone profile. We apply this procedure in an iterative way to polish all the generic timezone profiles.

4 Results of the Timezone Geolocation

We used our methodology to geolocate some of the most important Dark Web real forums. First, we collected information from several blogs on Tor and on the Dark Web. The Dark Web is large and very dynamic, therefore to test our findings we selected five forums amongst the best known and popular ones. Two of these are of known origin: The first, CRD Club, is mostly in Russian, whereas the second one, Italian DarkNet Community (IDC) is the forum of the homonymous Italian marketplace in the Dark Web. We use these first forums to validate and confirm our methodology, and then apply it to other 3 DarkWeb forums.

The experiments proceed in a similar way for all the forums. First, we sign up in the forum and write a post in the "Welcome" or "Spam" thread to calculate the offset between the server time (the one on the post) and UTC. In some forums the timestamp of the posts is accurate and already in UTC. In some other cases the timestamp does not specify the time zone and we might think that this information alone can uncover the location of the server (but not of the crowd of the forum). Of course, this is not the case since the timestamp can be deliberately shifted. In all cases, once the offset from UTC is known we can collect the timestamps of the posts in a sound and consistent way. Lastly, we also checked that in all of the forums the posts appear with no delay. This has been confirmed for all of the five forums that we have investigated.

4.1 CRD Club and the Italian DarkNet Community

The first case study is a Russian forum called the CRD Club (http://crdclub4wraumez4.onion). It is divided in two macro sections, the first one written in Russian (Cyrillic script), while the other one is an international section written in English. On this forum users write about technology, hacking, gambling, online anonymity, credit card frauds and selling. There is also a subsection for job offers—for example people looking for specialists that can hack a bank account or open a "bank drop" (an account open on fraudulent credentials, often in a fiscal paradise). After our analysis, we can conclude that this forum consists of a technology oriented crowd. Of course, we expect that our methodology locates this crowd in the Russian speaking countries.

We retrieved from the CRD Club 209 active users with $14,809$ posts in Russian. First, we note that the profile of activity of the users of the forum, shown in Figure 4(a), is very similar to the generic profile based on the whole Twitter dataset (Figure 1(c)). This observation is confirmed by the high Pearson correlation of 0.93 between the two profiles. This result supports the conclusion that the users of the Dark Web have similar access pattern of the users of the standard Web and therefore that the Twitter generic profile can be a good fingerprint for hidden services too.

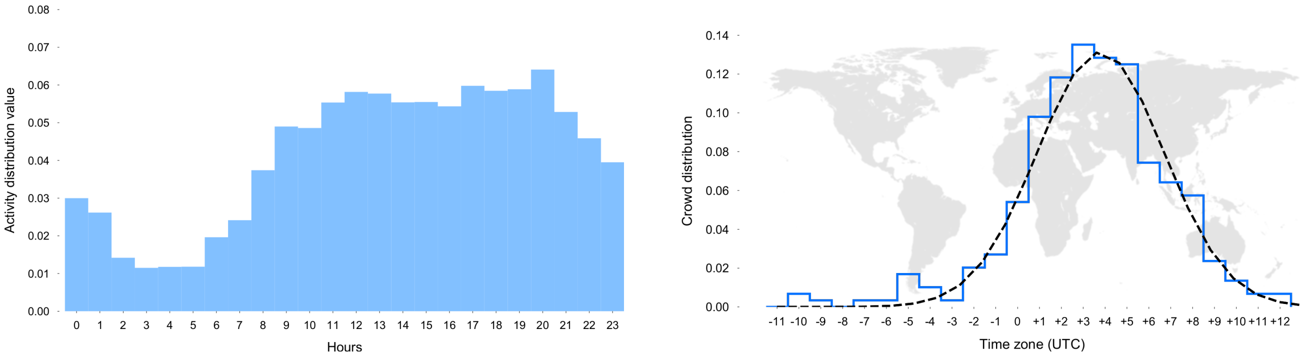
The results of our geographical classification is shown in Figure 4(b). As we can see, there is only one Gaussian component, with an average distance of 0.007 and a standard deviation of 0.006. This means that most of the crowd come from a specific geographical area. Moreover, the Gaussian mean falls between the UTC$+3$ (Bucharest, Moskow, Minsk) and the UTC $+$ 4 (Abu Dhabi, Tbilisi, Yerevan) time zones. We can note that a very large part of the population of the Russian speaking countries live exactly in these time zones.

We have done the same analysis for the Italian DarkNet Community (http://idcrldul6umarqwi.onion), a forum written in Italian and known to have an Italian crowd, with the same results that we omit due to space limitations.

4.2 Dream Market http://tmskhzavkycdupbr.onion

The Dream Market is the official forum of the Dream Market Marketplace. Most of the discussion is about the quality of goods and vendors in the marketplace, with a separate section to report scam vendors. It is an international forum, where English is the only language allowed. After polishing the data, we classified 189 users and $14,499$ posts.

In Figure 5(a) we show the results. As we can see, our methodology discovered two main Gaussian components with an average distance of 0.011 and a standard deviation respect to the crowd distribution of 0.008. The smallest component is centered in the UTC $-$ 6 time zone (Chicago, New Orleans, Mexico City) that is the American Mountain Time Zone. While the largest one is in the UTC $+$ 1 time zone (Berlin, Paris, Rome). We can note that the UTC $+$ 1 time zone, aside from Europe, covers also part of Africa



(a) Regional profile (UTC + 3).

(b) Gaussian distribution.

Fig. 4. The CRD Club Forum, http://crdclub4wraumez4.onion Russian Forum.

(Namibia, Zimbabwe, Nigeria, etc.), and, technically speaking, our methodology cannot rule out the fact that part of the crowd is from that part of the time zone, though this seems unlikely since Africa is less developed than Europe from a technological point of view. Rumors [10] suggest that the hidden service was under control by the Dutch police. One of the former administrators, OxyMonster, is French [11]. So, we believe that we can safely assume that the crowd of the forum classified in the UTC $+ 1$ time zone is mostly from Europe (though we cannot exclude that part of it, or in principle all of it, is from Africa).

4.3 The Majestic Garden

The Majestic Garden (http://bm26rwk32m7u7rec.onion) is a meeting point for people who share the passion for psychedelic experiences. It can be thought of as a virtual hippie commune in the Dark Web. In this forum the majority of topics are related to sharing experience related to drug assumption, in particular LSD and psychedelic mushrooms. There are also topics about selling and buying these substances or how-tos that can help you make them at home. In addition, there is a section dedicated to the literature on psychedelic and spiritual experiences. From this forum we classified $75,875$ posts from 638 active users. Their distribution is shown in Figure 5(b). We have again two main components, an average distance of 0.009 and a standard daviation of 0.011. The largest one is centered on UTC $- 6$ (Chicago, New Orleans, Mexico City), approximately in the Midwest belt. The mean of the second one falls into UTC$+1$ (Paris, Berlin, Rome). This is a mostly American forum.

4.4 Pedo Support Community

The users of this forum have a common interest in pedophilia (http://support26v5pvkg6.onion). As they say, the forum was born to share their experience far from a "world that does not understand who they really are". They are aware of the immorality and illegality of their interests and behavior—indeed in the forum it is possible to find some ethical discussion about their habits. Moreover, it is forbidden to share pedopornographic material in the forum. English is the mandatory language and is forbidden to disclose the country of the user. Lastly, some sections of the forum are hidden and access is allowed only to those that convince the administrators to be able to contribute to the discussions in a useful way. Of course, we have not done that. Therefore we have no data from that part of the forum.

After the cleaning step we classified 290 active users that wrote $44,876$ posts. In Figure 5(c) we show the distribution of users across the time zones. In this case we have three Gaussian components with a standard deviation of 0.012 and an average distance of 0.01. The highest one is centered between the UTC $- 8$ and the UTC $- 7$ (San Francisco, Los Angeles, Las Vegas) time zones. The second important component falls into the UTC $- 3$ time zone (Rio De Janeiro, Halifax, Sao Paulo). The last one is smaller and centered in the UTC $+ 4$ time zone (Yerevan, Tbilisi, Abu Dhabi).

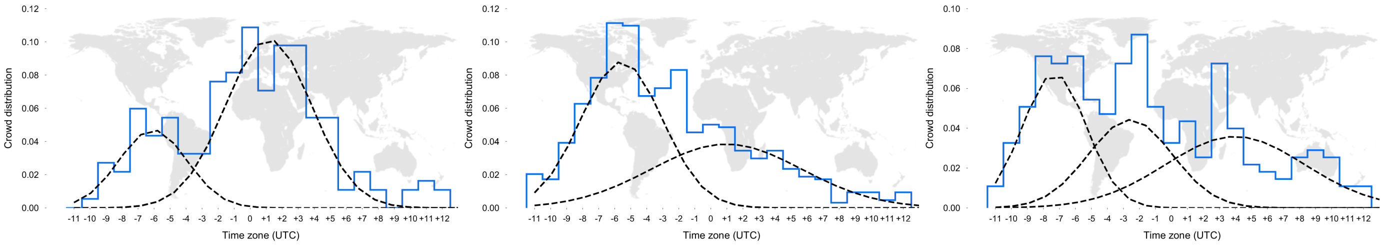
Differently from the other cases, in this forum we classified a component whose time zone, UTC-3 (Rio De Janeiro, Halifax, Sao Paulo), mostly covers countries in the southern hemisphere. The exception is Halifax, Canada, though its population is really small. So, intuition suggest that this part of the crowd lives in South America. To support this idea, we develop a methodology that we can use to indicate whether this crowd lives in the northern or southern hemisphere of the world. This is described in the next Section.

Lastly, for all five Dark Web forums under investigation, both the average and standard deviation of the point-to-point distance between the Gaussian curves and the crowd distributions shown in Table 2 are very low. Even more so when compared to the baseline values—those of the Malaysian distribution and its Gaussian fit shifted of 12 hours. This further supports our findings on these forums.

4.5 Discerning the Hemispheres

It is well known that daylight saving time consists in advancing clocks during summer. Usually, countries using daylight saving time adjust clocks forward one hour. The idea is to delay sunset during summer at the cost of a delayed sunrise to get more sunlight in the evening and save energy used for lighting. A simple observation is that this is done from (about) March to October in the countries of the northern hemisphere, while it is done from (about) October to February in the southern hemisphere. We can use this simple fact to understand if the people of the crowd lives in the northern or southern hemisphere.

We proceed in this way: If the profile of access to the Dark Web forum of a user in the period October–March is similar to the profile of the same user in the period March–October shifted one hour forward, we rule that the user



(a) Dream Market forum.

(b) The Majestic Garden.

(c) Pedo support community.

Fig. 5. Geographical classification of crowds in dark web forums.

lives in the northern hemisphere. Conversely, if the access profile in the period from October to March is similar to the profile of the same user in the period March to October shifted one hour backwards, we rule the user to live in the southern hemisphere. If we do not see any particular difference in the two periods, we assign the user to one of the countries that do not use daylight saving time without giving any information on the hemisphere. We use again the Earth Mover's Distance to measure similarity of access profiles.

To validate this procedure, we classified the five most active users in the datasets of United Kingdom, Germany, Italy, and Brazil. Note that all of these countries use daylight saving time (actually, in the case of Brazil, only the southern part of the country, the most populated, uses it). The 5 users in the dataset of United Kingdom as well the 5 in dataset of Germany and the 5 in the dataset of Italy, all of them, are classified as living in the northern hemisphere. The 5 users in the dataset of Brazil, all of them, are classified as living in the southern hemisphere. Therefore, we believe we can use this methodology with good confidence. We have done it for the Pedo Support Community, due to the controversial nature of the forum and the alleged origin from South America of a good part of the crowd. We limit our analysis to the 5 most active users of the forum, since those are the users for which we have a good number of posts. According to the analysis, 3 out of 5 of the most active users in the Pedo Support Community live in the southern hemisphere; the other 2 in the northern one. This result confirms our initial intuition that a good part of the crowd of the forum lives in South America. Actually in Southern Brazil or Paraguay, the only areas in the UTC $-$ 3 timezone of the southern hemisphere that do use daylight saving time.

5 Detecting the Native Languages of the Dark Web Crowds

Most posts in Dark Web forums are in English. Indeed, English is the lingua franca of the Web, and it is the language of choice of international communities. However, writings in English by non-native English speakers typically present distinctive patterns that are peculiar to the native language of the author. The idea is to use these peculiarities to uncover the origin (from a language point of view) of the crowd of the Dark Web forums. Towards our goal, we exploit Native Language Identification (NLI)—the task of automatically classifying the native language *L1* of a writer based on her messages written in another language *L2* [12]. In the Dark Web, our case, *L2* is English.

NLI works under the assumption that the native language *L1* is different from the language of the writing *L2*. But the Dark Web crowds are made of people from all around the world, including English native speakers. Therefore, we proceed in two steps: First, we tell apart native English speakers and speakers of English as a second language; then, we apply NLI to the latter group of users. Both phases are based on machine learning techniques. The datasets used in the learning process, however, have to be customized to the Dark Web users. Indeed, the standard language used on the Internet and, especially, in the Dark Web is so far from educated English that the datasets that can be found in the literature are not satisfactory for the task.

5.1 The Datasets

Distinguishing Dark Web English native from non-native speakers in an automatic way starting from their posts in English is not trivial. The first step is to find an appropriate dataset of text written in English by both types of users that can be used for the learning process. A dataset has labeling of the authors (users) in terms of mother tongue or, at least, nationality. We used the Wikipedia pages dataset [13], made of the user pages of Wikipedia contributors. These pages are typically divided into two sections: In the first, the user writes about herself; the second contains images and an informational box called *Babel*, a self-declared statement including the native language of the user and her proficiency in non-native languages. The dataset consists of $9,857$ users, $589,228$ comments, and 19 *L1* languages including English. This dataset proved to be effective in the learning process of the first step. The second step, identifying the mother tongue of the user writing in English in the Dark Web, is considerably harder. We had to add a dataset taken from the literature (TOEFL11) and integrate it with a new, custom one for our particular task.

TOEFL11 [14] was generated by sampling from a set of English essays written by TOEFL test takers. It is composed of $12,100$ essays with an average of 348 word tokens per essay. There are $1,100$ essays for each of the 11 native languages of the dataset (Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish). It is the dataset used in the NLI shared task in 2013 [12] and 2017 [15], and the most popular in NLI research.

The people that write on web forums have a different writing style compared to that of people taking an English exam. In fact, during the test, people take care of their writing and try to write in educated English. Instead, web writers have a more conversational style; they use abbreviations, idiomatic sentences typical of the Web, and bad language. Our goal is to detect the native language of these users, starting from posts written in this style. So, we need to integrate the TOEFL11 dataset with data similar to the writing style of Web users. Unfortunately, there was no dataset suitable for this purpose. Therefore, we built one from scratch making use of Reddit, a social news aggregation and discussion website and the sixth most visited web site in the world. Reddit posts and contents are organized by subject or by language (nation) based sub-communities into user-created boards called *Subreddits*. Through the Reddit APIs, we built the Reddit6 dataset that collects messages coming from 6 different *L1* languages: Italian, French, Spanish, German, Swedish, and Portuguese. It is composed of 6276 messages with an average of 320 words per post.

To generate the dataset, for each *L1* language (different from English) we identify as many native speakers of *L1* as possible. We do that by checking the subreddits written in *L1* (e.g., the subreddit `r/italy` for Italian and `r/de` for German) and tag the users of the subreddit as native speakers of *L1* if they post or comment in *L1*. In case of ambiguity—users writing in two languages different from English—we discard the user. Then, for each tagged user, we collect $1,000$ post in both $L1$ and English from all the subreddits in Reddit ($1,000$ posts is the maximum allowed by the Reddit APIs). We then apply a set of automatic operations: We discard users that look like bots, we remove duplicates due to crossposts, we discard messages that are too short or that are composed by a single sentence written multiple times, and we discard users with less than 10 messages in their native language. We also normalize URLs keeping only the hostname on the website. Then, we select only the messages written in English.

The detection of the language of the message in the above process is done through the Python library langdetect [16]. It is a Python porting of the Google Java library language-detection [17]. The library generates language profiles from Wikipedia and achieves a precision of over 99% for 55 languages. We also notice that some messages are written in more than one language. As an example, a message in English can contain a full phrase written in Spanish. To avoid these cases, we split each message into its sentences. Then, we apply the language detector for each sentence of the message. If the language detector classifies all the sentences with the same language, we label the message with the detected language. Otherwise, we discard the message.

After this automatic cleaning phase, we proceed manually to polish the dataset by doing the following operations: We discard messages with very long lists or messages with abbreviations or slang abuse. Lastly, we look for user statements about their native language, nationality, and other similar information that could allow us to confirm or contradict the labeling. Moreover, we consider the case of languages that are spoken in more than one country such as Spanish (e.g. Spain, Mexico, Argentina) or French (e.g. France, Quebec, Belgium). We are aware that there can be different inflections in different countries speaking the same language. Still, we believe that including these users in the dataset help classify correctly different dialects as the same language.

5.2 Native English Speaker Identification

5.2.1 Feature Selection and Classifiers

Feature selection is a key element in machine learning processes. We proceed in this way: We first tokenize the messages through the TweetTokenizer of the NLTK python package [18]. It can handle out of the box ASCII emoticons and replace character sequences of length greater than 3 with sequences of length 3, both very common on web messages. Then, we extract the features (see Tables 3 and 4):

Word *n*-grams A sequence of N contiguous tokenized words in a text. For our classifier we use n-grams of length 1 and 2 (respectively Word1 and Word2 in Tables 3 and 4).

Character *n*-grams A sub-sequence of N contiguous characters of a larger tokenized sequence. We use character n-grams of length from 3 up to 6 (Char3-6 in Tables 3 and 4).

Stem *n*-grams Stemming is the process of reducing inflected or derived words to their word stem or radix. After this pre-process, N stemmed contiguous words are taken together. We use stem n-gram of length 1 and 2 (respectively Stem1 and Stem2 in Tables 3 and 4).

We weight all features through the Term Frequency-Inverse Document Frequency (TF-IDF) computed as the number of times a word appears in a text (document) multiplied by the inverse fraction of the texts that contain the word. The TF-IDF gives less importance to the terms that tend to appear more often and in different texts.

To reduce the dimensionality of the features, we use the select K-Best with chi-square metric to select the K highest scoring features. As for the classifier, we use the same architecture described by Li et al. [19]. As done by Li et al., we used two classifiers: a Support Vector Machine (SVM) and a Multilayer Perceptron (MLP). SVM is a supervised learning algorithm that classifies by finding the hyperplane that maximizes the separation, also known as margin, between the two classes. Intuitively, given a set of data-points for which we don't have prior knowledge about their distribution, the optimal hyperplane is the one that divides the data and maximizes the distance between the two nearest data-points (margin) [20]. MLP is a feedforward neural network. It consists of a system of at least three layers of interconnected neurons or nodes. An MLP makes no prior assumptions on data distribution, and it can model highly non-linear functions. First, we build for each feature a linear SVM classifier with a L2 penalty, used to impose a loss to points that violate the margin. L2 is a penalty equal to the square of the magnitude of coefficients. Then we extract from each classifier the probabilities for each class and concatenate them. Finally we feed the MLP with the concatenated probabilities.

5.2.2 Results of the Learning Process

The learning process for the first step—native English speakers identification—is done by using 90% of the Wikipedia dataset as the training dataset. Table 3 reports on the F1-score for each feature and the final classifier. After

TABLE 3
Native English classification F1-score. Wikipedia dataset.

Feature Type	Total	K-best	F1Score
Word1	174, 646	15, 000	79.4%
Word2	1, 455, 954	50, 000	75.2%
Char3	73, 602	30, 000	81.8%
Char4	326, 145	30, 000	80.7%
Char5	979, 710	30, 000	78.9%
Stem1	125, 099	15, 000	80.5%
Stem2	1, 105, 393	30, 000	75.0%
Ensemble	—	14	82.2%

TABLE 4
Native language classification F1-score. Reddit6 and TOEFL11.

Feature Type	Total	Selected	F1Score
Word1	36, 167	15, 000	78.1%
Word2	251, 842	50, 000	65.2%
BHole	36, 167	15, 000	78.2%
Char3	15293	10, 000	70.5%
Char4	70, 938	30, 000	76.7%
Char5	199, 457	30, 000	79.2%
Char6	444, 581	30, 000	75.3%
Stem1	25, 572	15, 000	77.5%
Stem2	228535	30, 000	65.5%
Ensemble	− − −	99	81.6%

assess our classifier, we test it on the on the 10% of the Wikipedia dataset not used as the training set. The overall F1-score is 82.2%. More in details, our model performs very well in detecting not-native English speakers, classifying correctly 188 users out of 199. Also does it a very good job when detecting native English speakers since it classifies 140 users correctly out of 199. The overall F1-score is 82.2%. Moreover, we test our classifier on the Reddit6 dataset, whose users are not native English speakers. We obtain an F1-score of 93.5%, which confirms the high accuracy of our classifier in telling apart native and non-native English speakers.

5.3 Native Language Identification

5.3.1 Feature Selection and Classifiers

Feature selection for the second step—native language identification—is similar to the first step. However, the second step is considerably harder, and our experiments show that it is necessary to add additional features for the learning process. In particular, we add: **Hole *bi*-grams** A bi-gram of noncontinuous words. We use a window of size three and delete the one in the center to select the externals ones. We extract these features sliding the windows on the whole text. This feature makes the model stronger against typos and acronyms (BHole in Table 4).

Then, the techniques used to reduce the dimensionality of the features and the construction of the classifier is similar to what we have done for the first step.

5.3.2 Results of the Learning Process

Detecting the native language of web users is tricky for various reasons. A fundamental one was the lack of a dataset with web slang. Indeed, it is not easy to build a good dataset—web users are usually clustered in communities, and each community uses a different sub-slang, based on the topic. Moreover, each slang has its acronyms that sometimes can be the same but sometimes can differ (e.g. *lol* usually means "lots of laughs", in games communities it means "League of Legends"). For these reason, we built the Reddit6 dataset as described in Section 5.1. During the learning process, we split the Reddit6 dataset into two partitions: Training and test sets. The test set is composed of 10% of the samples randomly chosen for each class, while the training set is the remaining 90%. We perform three classes of experiments:

TABLE 5
Confusion Matrix for the Reddit6 test set

		Predicted			
		FRE	GER	ITA	SPA
Actual	FRE	65	5	6	7
	GER	2	86	2	4
	ITA	10	5	74	2
	SPA	5	6	4	44

TOEFL11 vs Reddit6. In the first experiment, we outline a baseline for our task. We start replicating the results described by Li et al. [19] (F1-score 86, 55%) using TOEFL11 as the training and test sets. Then, we apply the same classifier to the Reddit6 test set, achieving an F1-score of 56%. This gap in performance highlights the difference in the writing style between the TOEFL11 dataset and the web slang. However, the result is considerably above the random guess. This means that in the samples of TOEFL11 there are writing patterns that appear in Reddit as well.

Reddit6 vs Reddit6. We check the performance of the model training the classifier on Reddit6 and test it on its test partition. The F1-score achieved is 71.1%.

TOEFL11 and Reddit6 vs Reddit6. In our last experiment, since we have seen that there is valuable information in the TOEFL11 dataset, we merge the training sets of Reddit6 and TOEFL11, and test the classifier on the test set of Reddit6. The F1-score is 81.6%, considerably higher. This confirms that the work done to build a dataset specific for web slang has been fundamental. Table 5 shows the confusion matrix of this experiment for the Reddit6 test set.

TOEFL11 and Reddit6 vs TOEFL11. Finally, for the sake of completeness, we also test this last classifier against the TOEFL11 test set. In this case, our model achieves an F1-score of 85.8%. Even though our classifier has been tuned mainly around the problem of native language identification for web writers, it still does a good job when messages are written in educated English.

Now, we are ready to use our models in the Dark Web forums.

6 The NLI Methodology in the Dark Web

Native language identification of the users of Dark Web forums is a step forward in community profiling that integrates and supports the time-stamp methodology. We proceed in this way: For each forum, we first identify the native English speakers. Then, for those who speak English as a second language, we perform a further classification

TABLE 6
Native English Speakers in the Dark Web forums

Forum	Native	ESL
Dream Market	63	126
The Majestic Garden	439	199
Pedo Support Community	202	88
Total	704	413

TABLE 7
Tor Native Language Identification

Forum	FRE	GER	ITA	SPA	ARA
Dream Market	16	44	21	42	3
The Majestic Garden	21	64	57	56	1
Pedo Support Community	17	27	23	21	0
Total	54	135	101	119	4

to detect the native language. We apply this procedure to the Dream Market, The Majestic Garden, and the Pedo Support Community forums. We do not analyze the CRD Club forum and the IDC forum as most of the messages are in Russian and Italian.

6.1 Native English Identification in the Dark Web

The first step is to identify English native speakers. For each user, we select messages that are written in English and that are longer than 10 words: Our experiments show that 10 words are needed to get reliable results.

Table 6 shows the results of our classification. As we can see, both the Pedo Support Community and The Majestic Garden have a majority of English native speakers. Instead, the DreamMarket forum has a bigger community of non-native English speakers. These numbers seem to confirm the results of the time-zone methodology, which shows that the most significant Gaussian components for the Pedo Support Community and The Majestic Garden are in the American continent, whilst for the DreamMarket in Europe.

As a further experiment, we want to know where native English speaker are placed around the world. So, we integrate the results of the native English classification with the time-zone methodology. This way, we obtain a time-zone map where each user is located in her time-zone. Figure 6(a) shows the result of this operation for The Majestic Garden Forum. As we can see, most of the native English users are located in the American time-zones. Note that in the European and Asian zones, the majority of the users are non-native English speakers. For the Dream Market Forum, Figure 6(b), most of the users are non-native English speakers, and also in this case they are predominately in the European and Asian zones. Lastly, the crowd of the PedoSupport Community Forum, Figure 6(c), is mostly made of native English speakers from the American continent, except for the people in UTC -3 and -2, where the portion of non-native speakers is higher than the other forums.

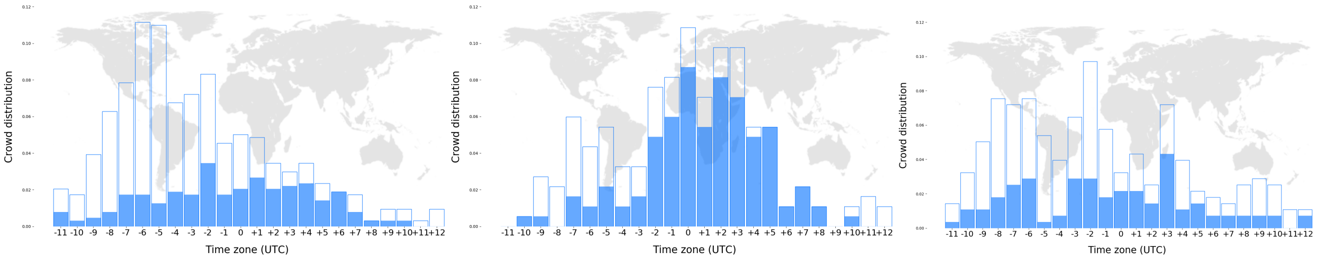
6.2 Native Language Identification in the Dark Web

To detect the native languages of the speakers of English as a second language, we use the classifier trained with both the Reddit6 and the TOEFL11 datasets. Before the analysis, we polish the messages as described in 5.1. Table 7 shows the results of our classification. As we can see, the German community seems to be the largest, though the Italian and the Spanish communities are very active too. The French community, instead, is the smallest in all the forums. Lastly, we found a handful of native Arabic speakers both in the Majestic Garden and in the Dream Market. To evaluate our results, we search the forums for users that explicitly declare their nationalities. At the end of our search, we found 26 users that we can use to validate our results. More in detail, we found 7 French users among which the Dream Market administrator OxyMonster [11], 9 Germans, 2 Italians, 7 Spanish and the Australian writer Eileen Ormsby also known on the web as OzFreelancer. After the classification process, OzFreelancer was correctly classified as a native English speaker and all the others as not native. So, we proceed to detect their native language. After the classification, we get that 20 out of 25 were correctly classified with their native language.

7 Related Work and Comparison

Native Language Identification. NLI has been largely studied in the past decades. Tomokiyo et al. [21] used a Naive Bayes system to identify native versus non native English speakers on the basis of POS and n-grams features. Al et al. [13] used Wikipedia to build a dataset to train a Linear SVM that reached an accuracy of 74.53%. The NLI task was first introduced in 2005 by Koppel et al. [22]. In this work they classified *L1* authors among five different languages (Czech, French, Bulgarian, Russian, Spanish) reaching an accuracy of 80.2%. A novel approach, based on the ensemble technique that fed the output of the first classifier as an input to the second one, was presented in [23]. Since then, the general trend in NLI was to use the ensemble method [12]. The work in [12] issued the first Native Language Identification Shared Task along with a new corpus, called TOEFL11 [14]. This dataset is still considered as one of the most popular in the community. Li et al. [19] built an ensemble of single features trained using SVM fed into a MLP for final result. For all their features they use a TF-IDF weighting approach reaching an F1 score of 86%.

Traffic Correlation Attacks. Many of the attacks to Tor in the literature are traffic correlation attacks to individual users. As Tor is a low-latency network and packet timing and size are not obfuscated, it is well known that an adversary able to observe both endpoints of a Tor circuit can de-anonymize the user [24], [25]. Bouer et al. [26] demonstrate that this kind of attack can be carried out in a quite efficient way. Entry nodes are chosen based on up-time and bandwidth rates reported by nodes, which are not verified by the Tor network. So, malicious nodes can maximize the likelihood to be chosen as entry nodes by reporting incorrect information about their up-time and bandwidth. Then, malicious nodes can also drop all circuits in which either of endpoints is non malicious. Hence the circuit must be rebuilt, and there is a new chance to build one with both endpoints under the control of the adversary. Correlation attacks can also be done at autonomous system



(a) The Majestic Garden forum. (b) The Dream Market forum. (c) The PedoSupport Community forum.

Fig. 6. Distribution of not native English speakers (blue) and native English speakers.

level. In 2009, Edman et al [27], using BGP (Border Gateway Protocol) historical routing data and simulating the path selection of Tor, show that a significant percentage of paths are vulnerable against an AS-level adversary. Nithyanand et al. [28], in 2013, found out that up to 40% of Tor circuits are vulnerable to the same adversary, 85% to a state-level adversary, and this value rises to 95% for states like China and Iran. Murdoch et al. [29] did similar work for an IXP-level adversary.

Network Manipulation Attacks. More recently, Sun et al. [30] introduce RAPTOR, a suite of three new attacks to de-anonymize individual users of Tor. In the first attack, they show that, instead of monitoring only one direction of the anonymous connections, an AS-level attacker can exploit the asymmetric nature of Internet routing. In this way, a malicious observer can increase the chance to observe at least one direction of the connections and use TCP headers in order to correlate them. In the second attack, they show that exploiting both the asymmetric correlation and the BGP churn a long term passive AS-level adversary can increase its surveillance capabilities by up to 50%. The last attack is an active one, where the adversary AS manipulates inter-domain routing by advertising incorrect BGP control messages. In this kind of attack the adversary can observe only one of the two connection endpoints and he launches a BPG hijacking attack against the not controlled endpoint in order to route the traffic into the malicious AS, allowing an asymmetric traffic correlation attack.

Website Fingerprinting. Other works are based on website fingerprinting [31], [32], [33]. With this approach the adversary builds a database of network traces of users who visit a set of websites—sequence of packets, size of packets in the sequence, inter-packet intervals. The adversary can do that in several ways, even by deploying his own users. Then, these traces are used to train a classifier. If the adversary can monitor the Tor traffic between a target user and the guard, he can use the classifier to learn what website is being visited by the victim. First results, despite the good accuracy score achieved in a controlled environment, show that in a open-world scenario the success of the attack is significantly lower [34]. Kwon et al. [35] used the same ideas, but this time monitoring only Tor circuits involved in a communication with the hidden services instead the whole Tor traffic. This adjustment greatly reduces the amount of connections to be monitored and makes the fingerprint attack feasible. In 2017, Overdorf et al. [36] repeat previous experiments with a largest dataset of .onion services, achieving an average accuracy of 80%. Moreover, they show that smaller and dynamic sites are the hardest to identify, larger and more static sites instead are more subject to this kind of attack.

Information Leak. Other attacks exploit information leaks from applications that were not intended to work over Tor. For example, Biryukov et al. [37] target Bitcoin users. They show that using Bitcoin through Tor exposes the users to a man-in-the-middle attack that can, over multiple transactions, identify the victim. Exploiting the anti-DOS protection of Bitcoin, an attacker can cause the ban of all non-malicious peers that run the Bitcoin protocol over Tor, forcing the victim to use a malicious exit node as peer. Now, the attacker can store a cookie inside the target client in order to identify the victim in future connections. Manils et al. [38] demonstrate that just the use of BitTorrent alone can jeopardize the anonymity of the user. In fact, in some case a BitTorrent tracker responding to a client query can disclose its IP address. So, an attacker that monitor the Exit node is able to unveil the identity of the client. Using a more sophisticated technique, a malicious tracker sends to the client a manipulated list of peers to connect the victim to a malicious peer and retrieve its IP.

Comparison with our work. In all the above mentioned works the attack targets the anonymity of individual users. In our work we uncover the geographical origin of a crowd of a Dark Web site, attacking the anonymity of the group instead of the anonymity of the single individual. Most of the previous works de-anonymize the users through correlation attacks or manipulating the network, where attackers need to control a portion of the network or monitor the traffic. These techniques are extremely effective to de-anonymize the users. It is possible to obtain even the users' IP. The drawback is that being able to monitor part of the network is extremely difficult and expensive, only service providers or government agencies are in a position to do it. And, even in that case, these techniques can only de-anonymize the individual users of that part of the network, not the whole crowd of the Dark Web forum.

Other works focus on information leak. While our methodologies can target only users that write on hidden services, with information leak is possible to target only users that use a particular software or a specific version of the software. The type of information that can be discovered depends on the specific flaw. Lastly, the goal of website fingerprint attack is to get information on the sites the user visits and not information about the user himself.

In all the attacks mentioned above, the attacker must be able to manipulate or monitor the network, at least an entry node. Moreover, the users under attack must be active during the attack. Our methodologies require only to gather and analyze the data in the Dark Web, and it can be carried out by anyone with a standard computer and Internet connection. Our methodology does not require that the user is still active on the network; in fact, the data we use are stored by the services under attack. Finally, our technique can be used in conjunction with the other methodologies by authorities to perform preliminary analysis, for example, to select the region or the AS to monitor.

To the best of our knowledge, this work is the first one that attacks anonymity by exploiting the collective behaviour of a crowd instead of technical weaknesses of the network or the protocols used in the Dark Web.

8 Discussion

No timestamp on posts Timestamps are always shown in the Dark Web forums under investigation. However, the forum might remove them to protect the time of access of the anonymous user. This is actually not stopping our methodology—it is enough to monitor the forum, see when posts are made and timestamp them ourselves. The process is slightly trickier than just creating a dump of all previous posts, as we have done in this work. One might need to monitor a sufficiently large number of days, depending on the frequency of the posts, in order to collect 30 post per user or more necessary to build meaningful profiles. Nonetheless, the methodology presented in this paper can still successfully be applied.

Forum shows and timestamps posts with random delay This is possible. But, to be effective, the random delay must be of at least a few hours reducing considerably the forum usability. So, many users could just move to other forums.

What if users coordinate to deliberately post with a daily activity profile of a different region? We assumed that people are not under the control of an adversary. Indeed, coordinating the behavior of hundreds of anonymous users can be very hard. Moreover, if anonymous users are forced to wake up in the night to make a post, most probably they don't, and they either leave the forum or keep behaving normally. **What if users intentionally post with personal information of different regions?** In our work, we used posts with statements about the region of the user to build the ground truth necessary to evaluate the performance of the de-anonymization. Of course, the users may have lied, but we believe that it is unlikely. We collected the messages from subreddits about non-critical topics, where users have little motivation to lie. Moreover, to be considered as a native speaker of a language, the user must have written at least ten messages in the language. The information about the user's region is used only to double-check. So, we can safely assume that if users with this behavior exist, they are too few to have an impact on the learning process.

Use of Adversarial Stylometry Users can use adversarial stylometry techniques, adjusting their writing style or using machine translation systems. It is known that adjusting the writing style is difficult but possible against the authorship attribution attack, the process of determining the writer of a document [39]. However, adjusting the writing style requires a continuous effort that one or few very motivated users can do, but it seems unlikely that a large crowd of a forum has the dedication to do it consistently.

9 Ethical Considerations

In this work we analyzed $1,378$ anonymous users of forums in the Dark Web. While doing so, we gathered $151,770$ posts from five different hidden services. The data collected from the Tor forums was encrypted and stored for a limited amount of time in our servers. It was not shared directly nor placed on platforms from where it could be downloaded. Consequently, and accordingly to the policy of our IRB, we did not need any explicit authorization to perform our experiments. Our work is compliant to the Tor research safety guidelines [40]. We believe that the Tor community can benefit from this research, that sheds light on important issues related to the privacy of Dark Web users.

10 Conclusion

In this paper we focused on geolocating crowds on the Dark Web into the time zones and countries of the World. The approach, that we believe to be unique in its kind, does not use traffic analysis or protocol-related breaches, unlike previous work. The fundamental idea is to build reliable profiles of posting activity on online forums, then, to match Dark Web crowd profiles to those of known regions, and to integrate these results with Native Language Identification. Our approach works well with crowds of users coming from a single country and many different countries. Further, it can be used to discover more fine-grained information on the crowds. An example is that of the most active users of the Pedo Support Community Forum in the Dark Web. We found out that an important part of the forum crowd comes from a region that covers Southern Brazil and Paraguay.

We believe that the methodologies presented in this paper lay down the foundations to shed light on the Dark Web and the multitude of its services from a sociological point of view. At the same time, our techniques can be particularly valuable to authorities performing ongoing investigation and geolocation of users engaged in illicit, cyber-criminal, or terrorism related activities in the Dark Web.

References

[1] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The second generation onion router," in *Proc. of USENIX Security*, 2004.

[2] zzz and L. Schimmer, "Peer profiling and selection in the i2p anonymous network," in *PET-CON 2009.1*, 2009.

[3] I. Clarke, S. G. Miller, T. W. Hong, O. Sandberg, and B. Wiley, "Protecting free expression online with freenet," *IEEE Internet Comput.*, 2002.

[4] M. Kihl, R. Larsson, N. Unnervik, J. Haberkamm, A. Arvidsson, and A. Aurelius, "Analysis of facebook content demand patterns," in *Proc. of IEEE SaCoNeT*, 2014.

[5] A. Arvidsson, M. Du, A. Aurelius, and M. Kihl, "Analysis of user demand patterns and locality for youtube traffic," in *Proc. of IEEE ITC*, 2013.

[6] A. Team. (2016) Archive team: The twitter stream grab. Accessed on 2017-06-12. [Online]. Available: https://archive.org/details/twitterstream.

[7] H. F. K, "The distribution of a product from several sources to numerous localities," *Studies in Applied Mathematics*, 1941.

[8] C. M. Bishop, *Pattern recognition and machine learning*. Springer-Verlag New York, 2006.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[10] Mashable Inc. (2017) Buyers fear police have secretly seized the last big dark web market standing. [Online]. Available: https://mashable.com/2017/07/21/dark-web-marketplace-drugs-dream-market

[11] The Guardian. (2017) Trip to world beard competition ends in arrest for alleged dark web drug dealer. [Online]. Available: https://www.theguardian.com/us-news/2017/sep/28/world-beard-moustache-competition-drug-dealer

[12] J. Tetreault, D. Blanchard, and A. Cahill, “A report on the first native language identification shared task,” in *Proc. of NAACL BEA*, 2013.

[13] Y. Chen, R. Al-Rfou, and Y. Choi, “Detecting english writing styles for non native speakers,” *arXiv preprint arXiv:1704.07441*, 2017.

[14] D. Blanchard, J. Tetreault, D. Higgins, A. Cahill, and M. Chodorow, “Toefl11: A corpus of non-native english,” *ETS Research Report Series*, 2013.

[15] S. Malmasi, K. Evanini, A. Cahill, J. Tetreault, R. Pugh, C. Hamill, D. Napolitano, and Y. Qian, “A report on the 2017 native language identification shared task,” in *Proc. of NAACL BEA*, 2017.

[16] M. M. Danilak. langdetect 1.0.7. Accessed: 2018-09-28. [Online]. Available: https://pypi.org/project/langdetect/

[17] N. Shuyo, “Language detection library for java,” 2010. [Online]. Available: http://code.google.com/p/language-detection/

[18] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O’Reilly Media, Inc.", 2009.

[19] W. Li and L. Zou, “Classifier stacking for native language identification,” in *Proc. of NAACL BEA*, 2017.

[20] V. N. Vapnik, “An overview of statistical learning theory,” *IEEE TNN*, 1999.

[21] L. M. Tomokiyo and R. Jones, “You’re not from’round here, are you?: naive bayes detection of non-native utterance text,” in *Proc. of NAACL*, 2001.

[22] M. Koppel, J. Schler, and K. Zigdon, “Automatically determining an anonymous authors’s native language,” in *Proc. of IEEE ISI*, 2005.

[23] J. Tetreault, D. Blanchard, A. Cahill, and M. Chodorow, “Native tongues, lost and found: Resources and empirical evaluations in native language identification,” *Proc. of COLING*, 2012.

[24] P. Syversonl, G. Tsudik, M. Reed, and C. Landwehr, “Towards an analysis of onion routing security,” in *Proc. of PETS*, 2001.

[25] V. Shmatikov and M. H. Wang, “Timing analysis in low-latency mix networks: Attacks and defenses,” in *Proc. of ESORICS*, 2006.

[26] K. Bauer, D. McCoy, D. Grunwald, T. Kohnoi, and D. Sicker, “Low-resource routing attacks against tor,” in *Proc. of ACM WPES*, 2007.

[27] M. Edman and P. Syverson, “As-awareness in tor path selection,” in *Proc. of ACM CCS*, 2009.

[28] R. Nithyanand, O. Starov, A. Zair, P. Gill, and M. Schapira, “Measuring and mitigating as-level adversaries against tor,” in *Proc. of NDSS*, 2016.

[29] S. J. Murdoch and P. Zieliński, “Sampled traffic analysis by internet-exchange-level adversaries,” in *Proc. of PETS*, 2007.

[30] Y. Sun, A. Edmundson, L. Vanbever, O. Li, J. Rexford, M. Chiang, and P. Mittal, “Raptor: Routing attacks on privacy in tor,” in *Proc. of USENIX Security*, 2015.

[31] D. Herrmann, R. Wendolsky, and H. Federrath, “Website fingerprinting: attacking popular privacy enhancing technologies with the multinomial naïve-bayes classifier,” in *Proc. of ACM CCSW*, 2009.

[32] A. Panchenko, L. Niessen, A. Zinnen, and T. Engel, “Website fingerprinting in onion routing based anonymization networks,” in *Proc. of ACM WPES*, 2011.

[33] T. Wang and I. Goldberg, “Improved website fingerprinting on tor,” in *Proc. of ACM WPES*, 2013.

[34] M. Juarez, S. Afroz, G. Acar, C. Diaz, and R. Greenstadt, “A critical evaluation of website fingerprinting attacks,” in *Proc. of ACM CCS*, 2014.

[35] A. Kwon, “Circuit fingerprinting attacks: Passive deanonymization of tor hidden services,” Ph.D. dissertation, Massachusetts Institute of Technology, 2015.

[36] R. Overdorf, M. Juarez, G. Acar, R. Greenstadt, and C. Diaz, “How unique is your. onion?: An analysis of the fingerprintability of tor onion services,” in *Proc. of ACM CCS*, 2017, pp. 2021–2036.

[37] A. Biryukov and I. Pustogarov, “Bitcoin over tor isn’t a good idea,” in *Proc. of IEEE S&P*, 2015.

[38] P. Manils, C. Abdelberri, S. L. Blond, M. A. Kaafar, C. Castelluccia, A. Legout, and W. Dabbous, “Compromising tor anonymity exploiting p2p information leakage,” *arXiv preprint arXiv:1004.1461*, 2010.

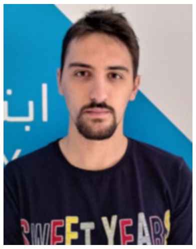
[39] M. Brennan, S. Afroz, and R. Greenstadt, “Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity,” *ACM TISSEC*, 2012.

[40] Tor Project. Tor Research Safety Board. Accessed: 2018-05-01. [Online]. Available: https://research.torproject.org/safetyboard.html

Massimo La Morgia is a Postdoctoral Researcher at the Computer Science Department of Sapienza University of Rome, Italy. He obtained the Laurea Degree in Computer Science, summa cum laude, and and the Ph.D in Computer Science from Sapienza University of Rome, respectively in 2014 and 2019. In 2014 and 2015 he won a research scholarship with the Computer Science Department of Sapienza University. His research interests include computer systems, security and privacy. He won the the ‘Avvio alla Ricerca 2017’ award from Sapienza University.

Alessandro Mei is a full professor and Head of the Computer Science Department of Sapienza University of Rome, Italy. He received the laurea degree in Computer Science summa cum laude from the University of Pisa, Italy, in 1994. He is a member of the ACM and the IEEE, a past associate editor of the IEEE Transactions on Computers (2005-2009), and the general chair of IEEE IPDPS 2009, Rome, Italy. Alessandro Mei was a Marie Curie Fellow in 2010–2012.

Eugenio Nerio Nemmi is a Ph.D student at the Computer Science Department of Sapienza University of Rome. He obtained the M.Sc. Degree, summa cum laude, from Sapienza University in 2018. From 2016 to 2018 he worked with the Computer Science department of Sapienza University, supported by a research scholarship. His activities along this collaboration focused on research in the areas of machine learning, security and privacy.



Simone Raponi is currently a 2nd year Ph.D Student of Computer Science and Engineering at Hamad Bin Khalifa University in Doha, Qatar. He is a teaching assistant of Cybersecurity and an active member of the HBKU Cyber-Security Research Innovation Lab. He received both his Bachelor and his Master’s Degree with honors in Computer Science at Sapienza University of Rome, Italy, working on topics strictly related to applied Security and Privacy. His major research interests include Cybersecurity, Privacy, Machine Learning and Artificial Intelligence.

Julinda Stefa is an Associate Professor at the Computer Science Department of Sapienza University of Rome, Italy. She received the Laurea degree in Computer Science, summa cum laude, and the PhD in Computer Science from Sapienza University of Rome respectively in July 2006 and February 2010.Her research interests include computer systems and network security and parallel and distributed systems. She is the co-General Chair of HotOS 2019. She was the recipient of the Working Capital PNI Research Grant by Telecom Italia (30 winners out of 2138) and of the Best Demo Award of IEEE INFOCOM 2013 and IEEE SECON 2013.