The Geography of Online Dating Fraud

Matthew Edwards*, Guillermo Suarez-Tangil†, Claudia Peersman*
Gianluca Stringhini†, Awais Rashid*, Monica Whitty‡
*Cyber Security Group, Department of Computer Science, University of Bristol, UK
†Information Security Group, Department of Computer Science, University College London, UK
‡Cyber Security Centre, WMG, University of Warwick, UK

***Abstract*—This paper presents an analysis of online dating fraud's geography. Working with real romance scammer dating profiles collected from both proxied and direct connections, we analyse geographic patterns in the targeting and distinct characteristics of dating fraud from different countries, revealing several strong markers indicative of particular national origins having distinctive approaches to romance scamming. We augment IP geolocation information with other evidence about the dating profiles. By analysing the resource overlap between scam profiles, we discover that up to 11% of profiles created from proxied connections could be assigned a different national origin on the basis of text or images shared with profiles from direct connections. Our methods allow for improved understanding of the origins of dating fraud, beyond only direct geolocation of IP addresses, with patterns and resource sharing revealing approximate location information which could be used to target prevention campaigns.**

I. Introduction

The online romance scam is one the most prevalent forms of mass-marketing fraud in many Western countries. False dating profiles are created by scammers as a prelude to a sustained false romance, during which the target is repeatedly defrauded of large sums of money. The impact on victims in terms of both monetary loss and emotional harm can be substantial. However, technical analysis of the methods used by these scammers remains sparse, with few quantitative analyses of attacks and attackers.

Previous work has explored victim understanding of the scam process in interview settings [1], text reuse in romance scammer approaches via Craigslist [2] and strategies deployed in an anonymous Chinese dating site [3]. A major unaddressed hurdle for combatting this fraud is understanding its true global origins, as misrepresentation of location is common. Uncertainty about location and international legal obstacles can hinder investigation and prosecution.

The locations scammers give in their profile are typically regarded as being as false as the profile picture, calculated to attract the interest of their targets [1]. Dating sites record the IP addresses used by scammers in creating and accessing their profiles, and may compare those addresses to blacklists or use the IP geolocation (especially when compared to the profile's declared location) to inform a judgement about the likelihood that a profile is genuine. In response, most scam profile authors make use of web proxies to disguise their IP address connection information, and so they appear to be using a connection from the location given in their profile information. Dating sites are predictably countering by banning access to their site via known web proxies and similarly allocated IP blocks. There are however limitations to the effectiveness of these countermeasures, with privately hosted or intentionally disguised proxies escaping the checks of proxy listing services.

The real location, even at a national level, of the creators of the scam profiles is of interest both to law enforcement and for other preventative efforts – not only for the purpose of identifying that a given profile is a scam, but for following up with appropriate countermeasures once a significant origin of scams has been identified (e.g., contacting local law enforcement, funding targeted preventative campaigns). This paper is the first study we know of to address this topic.

In this paper, we use a dataset of real online dating scam profiles which includes profiles created via both proxied and direct connections. We set out to answer the following research questions:

- **Where does dating fraud come from?** What does IP geolocation evidence tell us about the origins of profiles created via direct connections, and how does this connect to the locations given in the profiles?
- **Do profile elements get reused internationally?** Does reuse suggest different origins for dating profiles? Can we complement IP geolocation by examining profile elements being reused between unproxied and proxied connections?
- **Does dating fraud from different regions present different characteristics?** Do countries tend towards certain forms of romance scam in a distinctive manner?

In Section II below, we describe the available data, and note its limitations. In Section III below, we outline the significant origin countries within the SOURCE dataset, and the national locations those profiles present. In Section IV we look at text and images being shared between romance scam profiles, and what these patterns suggest about the PROXY dataset. In Section V, we examine the major scam origin nations to identify patterns in other elements of the profiles, before concluding in Section VI with a discussion of the policy implications of this analysis.

II. Data Source

The data used in this paper comes from a public online dating scamlist maintained at scamdigger.com, which offers up romance scammer profile data for public awareness. An exhaustive collection of the 5,402 scam profile instances, as collected during March 2017, was examined with respect to two sources of geographic information:

1) The location given in the scammer dating profile information.
2) The IP address used to create the profile, as reported by the dating site.

Other profile elements of note include the age, gender, occupation, marital status and self-description, which are analysed in detail in related work. Of the two sources of geographic information, the former was recorded as a string, often specifying location to a city level. This was geocoded to lat/lon coordinates and a standard format through queries to the Open Street Map's `Nominatim` service[1]. For the sake of brevity, the locations given in profiles are referred to as the *presented* locations.

The IP address information was mapped to a location through the use of a geolocation service [2], providing both coordinates and structured address information. Some 368 records contained no IP address information and were excluded, leaving 5,194 profile instances. Of the IP addresses used, many (67.9%) have been identified as known web proxies or VPN end-points by the dating site, raising doubts about the reliability of the inferred geographic location. For this purpose, we separate the data into the SOURCE (i.e., un-proxied users) and PROXY (i.e., proxied users) subsets, of 1,666 and 3,528 profiles respectively. It is possible that IP addresses from the SOURCE dataset are in fact unknown proxies, perhaps shared secretly amongst criminals, and similarly, it is possible that PROXY users are only masking their specific connection information rather than their national origins. We address these possibilities below as they touch upon our results.

Some important limitations of the data source must be acknowledged as context for our analysis. Firstly, the scamdigger.com site is primarily a scam-list for profiles submitted to a particular dating site, datingnmore.com, which reviews submitted profiles with particular focus on online dating fraud, and lists those identified as scammers either at registration or after interaction with members. The profiles presented are thus those of scammers that attempt to target this particular dating site, which may be a source of unknown bias. As with almost all criminal data analysis, these are also those dating fraud profiles from scammers who have been identified or caught, and it is possible that they are not representative of a more skillful subpopulation, which could also be geographically biased. The former issue could be explored further through comparison with statistics from other dating sites, where they can be persuaded to release this information. The latter is an inherent limitation of criminological data.

III. Geographic Origins of Dating Fraud

Table I lists the significant origin countries for the SOURCE dataset. The largest single origin by far was Nigeria, at over 30% of the dataset. West Africa in general accounts for over 50% of the SOURCE locations. These proportions closely match previous observations of the national origins of advance-fee fraud, as determined by email header IP addresses [4], [5], suggesting potential commonality between these types of fraud. The next largest origins, Malaysia and South Africa, are also well-known for producing other forms of internet fraud. All of the listed nations score below 50 on the 2016 Corruption Perception Index [6], except for the United States and the United Kingdom, suggesting these may be unusual cases.

	Nation	Count	Proportion
1	Nigeria	488	0.302
2	Ghana	216	0.134
3	Malaysia	178	0.110
4	South Africa	140	0.087
5	United Kingdom	86	0.053
6	United States	57	0.035
7	Turkey	50	0.031
8	India	47	0.029
9	Togo	41	0.025
10	Senegal	40	0.025
11	Philippines	29	0.018
12	Ukraine	28	0.017
13	Russia	24	0.015
14	Ivory Coast	23	0.014
15	Kenya	22	0.014

TABLE I: The SOURCE countries for > 20 scam profiles

Figure 1 plots the major scam origins against their profile's presented location, as directional arrows weighted by volume of scams. The United States is the location most commonly presented in dating profiles, at 63% of the SOURCE dataset, followed by the UK (11%), Germany (3%) and Canada (2%). As presented locations are usually indicative of the victims' nationality, we can understand the data as reporting that residents of the US are the major target of romance scams, followed by those of other western nations.

Africa: Most African sources focus their attention on the major western targets reported above. A notable exception is a cluster of profiles from Ghana which appear to report their location accurately. This may be a simple reaction to a scam-detection methodology which uses mismatches between presented and IP-geolocated locations[3]; or could represent a more 'honest' scam format aimed at extracting funds through straight seduction. A similar but smaller group appears in South Africa. Other exceptions include a small cluster of profiles from South Africa and Ghana which present their location as Iraq and Afghanistan. These are classic "military scam" profiles, purporting to be members of the US military stationed overseas. A small number of Nigerian profiles present their location as Malaysia, for unclear reasons.

Europe: Almost all SOURCE profiles from the United Kingdom presented themselves as from the United States, with only 9% targeting the United Kingdom itself, despite this also being an internationally targeted location. Profiles originating in Turkey targeted the United States and Germany, in keeping with the international norm. Most interestingly, profiles from the Ukraine and Russia almost always presented their national location as consistent with their IP address. This marked deviation from the pattern of romance scams originating elsewhere highlights the distinctive nature of Russian and Ukrainian dating fraud.

[1] https://wiki.openstreetmap.org/wiki/Nominatim (March 2017)
[2] http://freegeoip.net (September 2017)
[3] Such a method is in use by the dating site operators

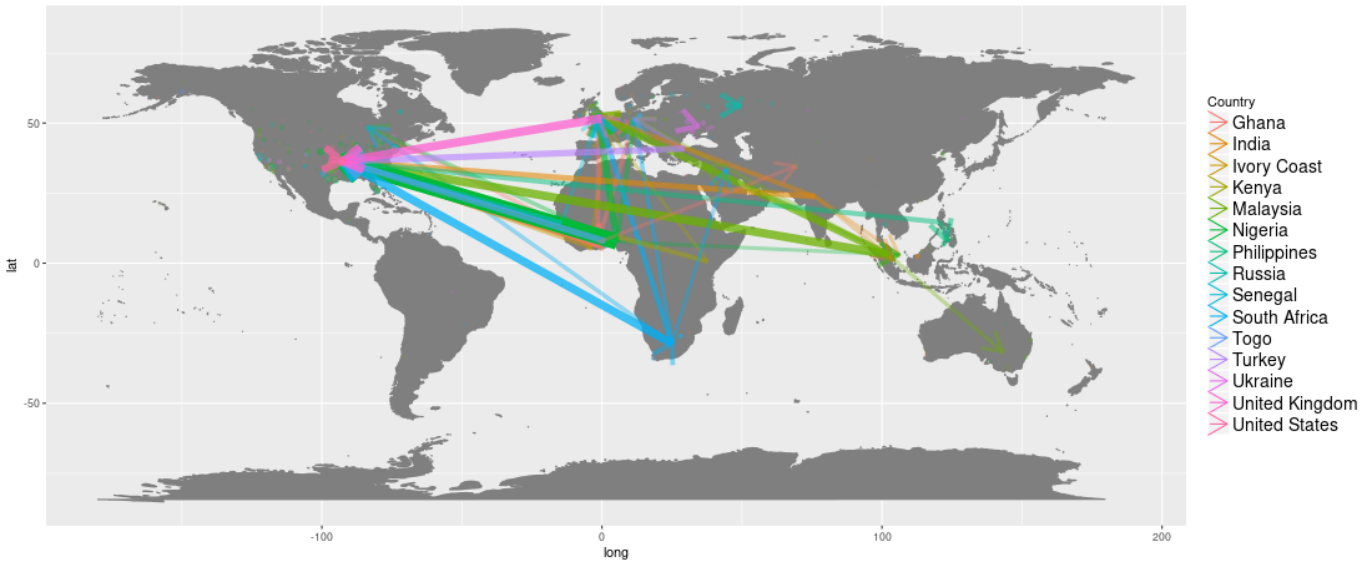


Fig. 1: The major paths from SOURCE IP addresses to the locations given in profiles

Asia: India follows the international norm in presenting profiles as from the United States and United Kingdom, although the ratio allocated to each is weighted more in favour of the United Kingdom (2:1 vs the 10:1 in West Africa), perhaps due to closer national ties. There are some small groups of Indian source IPs which present profiles in Singapore or Malaysia. Malaysian scammers also present profiles in the US and UK at the Indian 2:1 ratio, with small secondary clusters presenting from Malaysia and nearby Australia. Scammers in the Philippines split their presentation between the Philippines itself and the US, an unusual pattern that likely reflects the close links between the US and the Philippines.

United States: Almost all SOURCE profiles from the United States gave their location as within the United States. However, the most common presented state locations were New York and Texas, while the source addresses were mostly located in Arizona, California and Virginia, suggesting a degree of location misrepresentation within the nation or else imprecision of unknown proxying attempts.

IV. Augmenting Geolocation Evidence

As previously highlighted, SOURCE IP addresses are not necessarily accurate origins – they could be unknown proxies which escaped detection. While this is inherently an unknown factor, we can make use of certain additional evidence as an augmentation. For SOURCE IP information we can assess the likelihood of impersonations, and for the unknown PROXY subset's true locations we can examine the reuse of text and images with direct connections.

A. Probabilistic Assessment

We can first estimate the likelihood of this possibility by comparing the ratio of SOURCE and PROXY IPs for national locations. It is known that proxy lists will have a certain degree of error or incompleteness, which, under a base assumption that knowledge of proxies is affected similarly despite their location around the globe, means we are searching for an unknown threshold at which to discard the idea that certain origins are genuine – the rate of false negative error in these proxy lists. As we cannot be certain of this rate, no hard conclusions can be drawn from proxy ratios alone, but we can say that a large SOURCE:PROXY ratio is a signal carrying some information about the credibility of location information. Where the number of profiles with an unknown IP address is a small fraction of the number of known proxies for this location, we will regard these locations as suspect. Where this is not the case, we can be more confident that the IP address accurately reflects the origin of the scam profile.

Nation	PROXY	SOURCE:PROXY
United States	1949	0.03
United Kingdom	204	0.42
Russia	47	0.50
Ukraine	23	1.17
Philippines	11	2.42
Turkey	10	4.55
India	5	7.83
Kenya	1	11.00
Ivory Coast	0	23.00
Malaysia	5	29.67
South Africa	3	35.00
Nigeria	12	37.54
Senegal	0	40.00
Togo	0	41.00
Ghana	4	43.20

TABLE II: Ratio of suspected source IPs to known proxies by country

Table II presents this ratio for the major SOURCE countries. From this, we can say that we have the most reason to be suspicious of the validity of IP addresses situated in the United States, with the observed count of scam IP addresses not known to be proxies being a very small fraction of those from known proxies. We also know that the majority of the SOURCE

dataset from outside the US have presented their location as being in the US, attesting international effort at exactly this form of misinformation. Looking at temporal reporting information, we find that the proportion of SOURCE profiles in the US has been decreasing since 2013, suggestive of gradually improving proxy detection.

The UK is the next most suspect IP location, also attracting a large volume of SOURCE profiles as a falsely presented location, and with more PROXY than SOURCE IP addresses. However, scammers would have to be an order of magnitude more effective at masking their IP addresses as UK locations than as US locations, in order to explain the ratios of scam profiles generated by these IP addresses. It is notable that both SOURCE and PROXY profiles from UK IP addresses most often present themselves as located in the US. This suggests either that the UK supports a population of relatively security-conscious romance scammers targeting the US, or is acting as a significant staging ground for fraud from elsewhere directed at the US. Temporal information here also suggests a downward trend since a spike in 2014.

Russia and the Ukraine are also locations with a significant number of PROXY profiles, but here there is less reason to suspect the SOURCE IP addresses do not reflect the national origin of the scam. Unlike the US and UK, we do not see any significant number of other SOURCE profiles presenting Russia and the Ukraine as their location, and unlike the SOURCE profiles, most PROXY profiles from these locations present their location as the US. The reporting figures appear stable over the observed period. The few presented Russian and Ukrainian PROXY profiles may simply be scammers protecting their individual location and connection information, without interest in masking their national origins. Similarly, known proxies account for just over a quarter of the IP addresses from the Philippines, but there are few profiles traced from outside the country which purport to be located there, so there is little reason to suspect large-scale misrepresentation.

The remaining locations are only lightly populated by IP addresses from known proxies, and we may have confidence that these are genuine national origins of online dating fraud.

Some locations show up neither as significant SOURCE origins nor as presented locations in profiles, but only as transit points in the PROXY dataset. These are locations with significant proxy populations, but apparently of low appeal as targets for international dating fraud. All such profiles predominantly presented as located in the United States, with the proxy country being at best a distant second. Notable transit locations include the Netherlands, Switzerland, Sweden, France, Australia, Romania and Finland.

B. Profile Description Reuse

Previous work has shown that romance scammers engage in substantial reuse of certain profile elements to save on labour, using certain cached images and making use of textual "scripts" which can be copied and pasted with minimal editing [2]. We here seek to explore how these sharing patterns appear geographically. Understanding which sources are sharing resources can help identify cooperating criminals and similar scam types. Geographic clusters of resources can also be useful in identifying the true origins of profiles using proxies to hide their location.

Text reuse is common in scam profiles, with key chunks of text and expressions being observed across different unique profiles. To identify these overlaps, we first preprocessed the textual descriptions to standardise case and remove punctuation, and then used a *longest common substring* method to cluster texts. Any two texts which shared more than a threshold of 10 tokens (words) were considered to be part of the same cluster. By this method, 899 unique profiles could be assigned to a cluster, sharing text with at least one other profile[4]

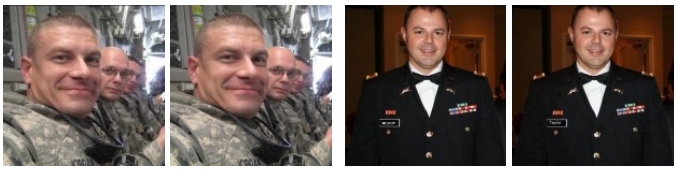
Location	Assigned
Nigeria	88
Ghana	56
Malaysia	41
Italy	11
South Africa	8
India	5
United Kingdom	5
Benin	4
Kenya	4
Philippines	4
Other	15

TABLE III: Inferred true locations of PROXY profiles

Looking first of all at reuse within the SOURCE subset, the greatest text reuse occurred within nations, with multiple unique profiles originating in Nigeria and South Africa sharing description text. The greatest international text reuse was between Nigeria and South Africa, with multiple profiles in each country sharing elements, and, interestingly, between Nigeria and the United States. Given the previous evidence that the SOURCE profiles in the United States may have been created through undetected proxies, we can take these Nigerian and South African scripts appearing in the US as further evidence of this under-detection. Similarly, scripts appearing in the United Kingdom suggest that there are undetected proxies amongst the SOURCE IP addresses from the UK. Text reuse within Africa and between Nigeria and to a lesser extent with all of Malaysia, India and Turkey, suggest a common approach to romance scamming in these nations. Notably, we see little to no direct text reuse from Russia, the Ukraine or the Philippines, either internally or externally, though it is worth noting that we have relatively few examples from these countries in comparison to the numbers from West Africa.

Turning to the PROXY dataset, we find that 241 (11%) share text with SOURCE profiles, meaning that their true location can be indirectly inferred. Table III reveals the results of assigning the majority national label for shared clusters. As well as adding significantly to the totals for the already-dominant West African and Malaysian scam origins, this inference also reveals a number of Italian scam profiles. Combining these discovered origins with the smaller number of Italian SOURCE profiles which enabled this inference, Italy would place 11th in Table I, with more profiles originating here than in Russia or the Ukraine.

[4]This number does not count variants of the same profile identified as such from the dataset, so these 899 reflect 28% of the dataset



(a) *2d...a8.* (b) *d4...9e.* (c) *6c...89.* (d) *15...bf.*

Fig. 2: Images reused by scammers in different profiles. Each sub-caption shows an excerpt of the hash of the image. Note that although certain images are perceptually equal, their hashes are different.

C. Profile Image Reuse

The use of images plays an important role in online dating sites. Scammers often reuse profile images that have been shown to attract vulnerable users in other locations. The military, the academic and the medical context are recurrently exploited [1]. Figure 2 shows four examples of image reuse. These images appear in different scammers' profiles. While some images are perceptually the same picture, their hashes are totally different. This is the case, for instance, of Figure 2a and 2b, where their hashes are *2da1883450f2b74357465d3031cfd2a8* and *d43c4519edc110c6a53dd10e40414e9e* respectively.

In our work, we use perceptual hashing to fingerprint images. This type of hashing extracts features from the images so that two images will have the same perceptual hash when features are similar. These hash functions can distinguish between dissimilar images, while being robust against different transformations and "attacks" [7]. For the purpose of this paper, we leveraged different perceptual hashing algorithms including the classic *perceptual hash* function—computed from the Discrete Cosine Transform (DCT) between the different frequency domains of the image—and *wavelet hashing*—using the Discrete Wavelet Transformation (DWT) [8]. Perceptual hashes within the dataset are compared in a pairwise manner using their Hamming distance, and then tested for equivalence based on a distance threshold, and manually verified to exclude false positives. We observe a total of 187 images which are perceptually equivalent, with some being reused across up to five different scam profiles.

Image clusters were then aggregated from perceptually equivalent images which were connected by being presented on the same profile page (our assumption being that these are attempts to portray the same subject, even if perceptually dissimilar in setting). There were a total of 183 profiles connected by 57 image clusters. Within the SOURCE subset, there were 45 profiles connected by 27 clusters of images.

Examining reuse within the SOURCE subset, images were predominantly shared between profiles created within Nigeria (14 internal connections to 4 external), Ghana (12 to 2), the UK (5 internal) and South Africa (5 internal). The external connections from Nigeria and Ghana were to Ghana, Nigeria, Benin, Kenya and Turkey. Though the numbers here are small, they fit with the more substantial body of text evidence showing resource sharing largely appearing to happen within nations in West Africa. These images might be copied from other scammers, or profiles in our dataset could have been created by the same scammer under an unresolved alias.

Turning to the PROXY subset, 19 image clusters in this data were connected to the SOURCE subset, allowing a total of 48 proxied profiles (1% of the subset) to be connected to profiles from unproxied connections. The major connected locations were Nigeria (22), Ghana (13), Togo (5) and the UK (4), with the majority of the PROXY profiles affected presenting a US location. Again, this is congruent with other evidence of a largely West African scammer population making use of proxied connections to present themselves as US citizens, with some hints of scammers also acting from within the UK.

V. Characterising Geographical Differences in Scam Profiles

A previous section has explored how the presented location in a scam profile can differ according to the actual location of its creator. Other profile elements may also vary geographically, according to the particular flavours of romance scam being employed in each location. In the section below, we examine how demographic characteristics are distributed according to the origin of scams from the SOURCE dataset.

We survey the demographic information—age, gender, occupation, ethnicity and marital status—for each major scam origin country[5] in the SOURCE dataset. Z-tests were performed for the age, gender, and topmost category of occupation, ethnicity and marital status, compared to the SOURCE population averages. Table IV presents the results, with statistically significant differences ($\alpha = 0.05$) highlighted in bold. Bonferroni correction was applied to adjust for multiple comparisons. Gender is presented as the proportion of males.

An immediate division can be drawn between countries which predominantly present male profiles (e.g., Nigeria, Malaysia, South Africa) and those which present mostly female profiles (e.g., the Philippines, Ukraine, Senegal). The age of scam profiles corresponds with their gender, with female scam profiles typically averaging around the age of 30, and male profiles averaging towards 50. The rates by which profiles declare themselves single also appear to be gender-biased, with female profiles being far less likely to use alternative statuses such as divorced or widowed. These would seem to correspond to very different top-level strategies of online dating fraud being pursued in different countries, with, presumably, different targets in mind.

Within nations presenting mostly male profiles, the strategies appear to be fairly similar. They all mostly report white ethnicities, and most frequently use military or engineering occupations. Two exceptions are India, where the all-male scam profiles mostly present themselves as 'businessmen', and Italy, where the profiles most commonly report professions in the real estate sector. Marital status provides the most interesting distinctions. It is clear that a heavy use of the 'widow' backstory is especially favoured by South African and Turkish scammers, also most evident in the profiles with

[5]Those in Table I, plus Italy, which is promoted to importance when considering text reuse evidence

Nation (SOURCE IP)	N	Age		Gender		Occupation			Ethnicity			Marital Status		
		$\bar{x}$	z	$\bar{x}$	z	x	$\bar{x}$	z	x	$\bar{x}$	z	x	$\bar{x}$	z
Nigeria	488	42.61	0.95	**0.73**	**4.13**	military	0.19	1.54	white	0.60	-1.38	single	0.47	-1.73
Ghana	216	40.01	-2.81	**0.46**	**-5.34**	military	0.22	2.07	white	0.68	1.65	**single**	**0.63**	**3.70**
Malaysia	178	**46.53**	**5.33**	**0.79**	**4.29**	**engineer**	**0.30**	**5.71**	white	0.60	-0.86	single	0.46	-1.17
South Africa	140	**48.61**	**6.95**	**0.81**	**4.35**	engineer	0.22	2.15	**white**	**0.77**	**3.57**	**widow**	**0.57**	**7.47**
UK	86	**46.15**	**3.38**	**0.93**	**5.64**	**military**	**0.33**	**4.09**	white	0.66	0.71	**divorce**	**0.33**	**5.51**
USA	57	**47.33**	**3.56**	0.84	3.21	**engineer**	**0.34**	**3.71**	white	0.61	-0.20	widow	0.30	0.18
Turkey	50	46.08	2.53	0.72	1.21	military	0.26	1.82	**white**	**0.86**	**3.48**	**widow**	**0.58**	**4.66**
India	47	42.62	0.30	**1.00**	**5.17**	**business**	**0.32**	**8.14**	white	0.62	-0.14	single	0.53	0.39
Togo	41	**39.44**	**-1.56**	**0.20**	**-5.89**	**military**	**0.37**	**3.52**	white	0.39	-3.20	single	0.68	2.34
Senegal	40	**33.98**	**-4.67**	**0.10**	**-7.07**	**student**	**0.57**	**11.23**	**black**	**0.38**	**7.75**	**single**	**0.88**	**4.81**
Philippines	29	**27.66**	**-7.07**	**0.03**	**-6.75**	**sales**	**0.50**	**14.85**	**mixed**	**0.48**	**7.81**	**single**	**0.97**	**5.14**
Ukraine	28	**29.15**	**-6.11**	**0.07**	**-6.23**	**academic**	**0.22**	**9.82**	**white**	**1.00**	**4.24**	**single**	**0.89**	**4.26**
Russia	24	**29.25**	**-5.72**	**0.08**	**-5.65**	**accounts**	**0.43**	**23.18**	**white**	**0.96**	**3.51**	single	0.79	2.94
Ivory Coast	23	36.52	-2.44	**0.30**	**-3.32**	**student**	**0.30**	**3.86**	**black**	**0.48**	**8.02**	single	0.65	1.48
Kenya	22	35.73	-2.72	0.45	-1.79	**self**	**0.32**	**3.28**	white	0.55	-0.82	single	0.64	1.30
Italy	19	39.37	-1.09	0.89	2.33	**realty**	**0.63**	**23.74**	white	0.89	2.55	**single**	**0.89**	**3.59**
SOURCE	*1666*	*42.13*	-	*0.64*	-	*military*	*0.17*	-	*white*	*0.63*	-	*single*	*0.50*	-

TABLE IV: Dominant demographic characteristics by origin country. Significant differences highlighted.

the (suspect) location in the USA. The UK features profiles unusually willing to make use of a 'divorced' status. These exceptions mark certain nations as following variant patterns from the "Nigerian" approach. The commonalities between countries could be attributed to larger-scale campaigns of location disguise of the same individuals, or an international criminal group or community following similar patterns of activity, perhaps actively sharing tactics.

Nations presenting mostly female profiles have more markedly distinct strategies. Profiles from Senegal and the Ivory Coast are most likely to state a black ethnicity, and to report being students. Profiles from Russia and the Ukraine are almost universally white, but may be distinguishable by their declared occupation, with accounting professions being particularly distinctive of Russian profiles. Profiles from Togo are notable for including a number of females reporting military occupation, whilst the Philippines often present distinctively as mixed-race, working in sales positions.

Ghana presents an unusual picture, with a distinctive bias towards a more mixed-gender approach to scamming. Further examination reveals that Ghanaian profiles represent a mix of two competing local approaches: male profiles following the preferred pattern from nearby Nigeria, with dominantly military occupations, whilst female profiles borrow from a tradition more akin to that in Senegal and the Ivory Coast, presenting mostly as students – though, interestingly, Ghanaian female profiles are still more likely to be white than black. Kenya also appears to represent a balanced gender mix of scam profiles, but the comparatively small number of profiles from there make it difficult to be confident about this pattern.

The language in profile descriptions also shows some regional characteristics. To analyse the variety of topic categories found in the profile descriptions, we used dictionary terms that are mapped to categories from the LIWC 2015 dictionary [9]. Normalised category frequencies were recorded for each profile description and grouped by country of origin.

As can be seen in Figure 3 our results showed that scammers based in Russia referred considerably more to personal concerns, such as work, leisure and especially religion, and that

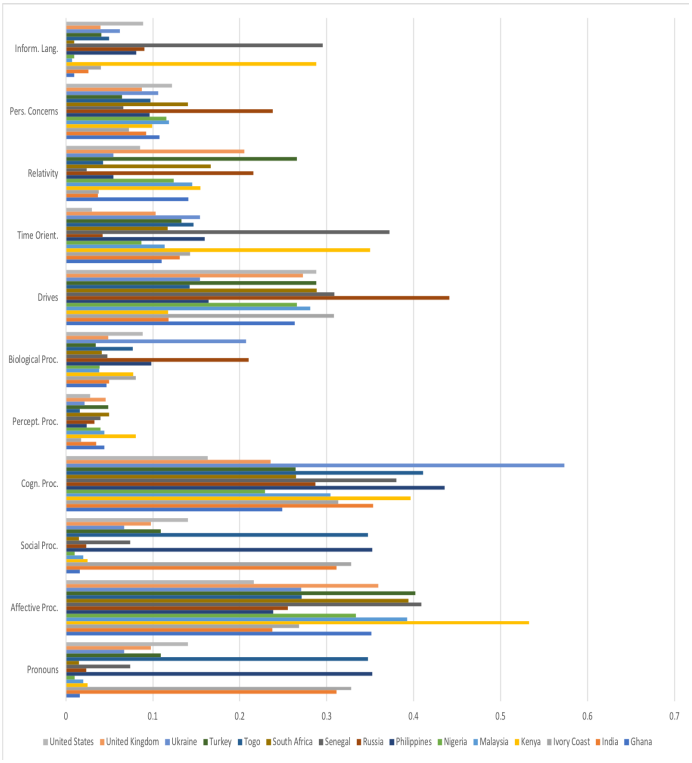


Fig. 3: The variety of topic categories [9] found in scammers' profile descriptions by origin country.

they tend to focus on their motives or drivers (e.g. affiliation, status, power), while scammers from Togo, the Philippines, the Ivory Coast and India used more words related to social processes (e.g. references to friendship and the use of personal pronouns such as "I", "you", "he", "she", "we", "they"). Interestingly, when analysing the number of references to cognitive processes, such as tentative language use (e.g. "maybe", "perhaps", etc.), discrepancies (e.g. "should", "would", etc.) and cognitive words (e.g. "know", "cause", etc.) the highest number of references was found among Ukrainian scammers. Finally, Kenya, and to a lesser extent, Senegal, stood out for

their language use linked to negative emotions, such as sadness and affective processes (e.g. "cry") and for displaying more informal language forms (including 'netspeak' features).

It appears that there are certainly regional characteristics of online dating fraud, which appear to be distinctive. Further work will discuss whether demographic and linguistic features such as these are robust enough to automatically identify the origins of scam profiles.

VI. Conclusion

We have provided an overview of the geography of online dating fraud, to the extent that our data source allows us to explore this topic. While most online dating fraud profiles present their location as a major location in the US or other Western country, their origins are mostly West African, Malaysian or South African, with Nigeria the largest single contributor. These may be the most important targets for efforts at preventing online dating fraud. Preventative and disruptive efforts, working with regional agencies in these locations, could have significant international impact.

Treating IP location information critically, we observe that profiles which appear to have been created from the United States can often share text or images with scam profiles being created from elsewhere. Text reuse indicators can suggest true locations for up to 11% of data coming from known proxies. This is only an initial analysis, but refining and applying this methodology using a database of known scam profiles could help capture new scam profiles reusing observed elements. Methods such as these, working from larger evidence-bases, could help form a technological countermeasure to increasing utilization of proxies amongst scammers, using the wider criminal population to help identify and locate the more careful elements.

At a national level, most countries producing romance scam profiles tend toward creating mostly male or mostly female profiles, suggestive of different cultures motivating the activity. With a few exceptions, mostly-male-profile countries follow a similar strategy with profile demographics, while mostly-female-profile countries appear to have more distinctive trademark approaches, which may be useful for investigators assessing the likelihood of particular origins. Further work is needed to determine how reliable presented characteristics alone can be for determining the origins of a given profile.

While a certain number of fraudulent dating profiles are seen to originate in the US, a number of other indicators suggest that many of these profiles come from undetected proxies for West African dating fraudsters, and the true rate for dating fraud originating in the US may be much lower than IP geolocation evidence alone would suggest.

The findings here might be taken up by law enforcement and government to guide global efforts at disruption of this particular crime. Whilst it comes as no surprise that these crimes emanate from West Africa, according to our data, the UK also have a share of fraudsters, and proxies across Europe are being used as transit points to target victims in the US. Law enforcement in Western countries, therefore, need to be concerned with both the crimes that enter the country from abroad and the criminals located within their own territory.

References

[1] M. T. Whitty, "The scammers' persuasive techniques model: Development of a stage model to explain the online dating romance scam," *British Journal of Criminology*, vol. 53, no. 4, pp. 665–684, 2013.

[2] T.-F. Yen and M. Jakobsson, "Case study: Romance scams," in *Understanding Social Engineering Based Scams*. Springer, 2016, pp. 103–113.

[3] J. Huang, G. Stringhini, and P. Yong, "Quit playing games with my heart: Understanding online dating scams," in *Proceedings of the International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2015, pp. 216–236.

[4] M. Kienpointner, "How to present fallacious messages persuasively: The case of the nigeria spam letters," *Considering Pragma-Dialectics*, pp. 161–173, 2006.

[5] E. Edelson, "The 419 scam: information warfare on the spam front and a proposal for local filtering," *Computers & Security*, vol. 22, no. 5, pp. 392–401, 2003.

[6] Transparency International. (2016) Corruption perceptions index. [Online]. Available: https://www.transparency.org/news/feature/corruption_perceptions_index_2016

[7] C. Zauner, M. Steinebach, and E. Hermann, "Rihamark: perceptual image hash benchmarking." in *Media Forensics and Security*, 2011, p. 78800X.

[8] V. Monga and B. L. Evans, "Perceptual image hashing via feature points: performance evaluation and tradeoffs," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3452–3465, 2006.

[9] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," Tech. Rep., 2015.