INACH Annual Report 2016-2017

By Tamas Berecz and Charlotte Devinat

Project Research - Report - Remove: Countering Cyber Hate Phenomena

INACH

International Network Against Cyber Hate

Supported by the Rights, Equality and Citizenship (REC) Programme of the European Union

Executive Foreword

This publication was written within the framework of the Research – Report – Remove: Countering Cyber Hate Phenomena project of the International Network Against Cyber Hate (INACH); funded by the European Commission Directorate-General for Justice and Consumers. The duration of the project is 2016-2017, and its aim is to study, document and report on online hate speech in a comparative and comprehensive way; and to establish structures for a transnational complaints system for instances of cyber hate.

Hate speech is intentional or unintentional public discriminatory and/or defamatory statements; intentional incitement to hatred and/or violence and/or segregation based on a person's or a group's real or perceived race, ethnicity, language, nationality, skin colour, religious beliefs or lack thereof, gender, gender identity, sex, sexual orientation, political beliefs, social status, property, birth, age, mental health, disability, disease.

This report was completed with the participation of the different members of the Network and partners in the project, namely the Zivilcourage und Anti-Rassismus-Arbeit (ZARA) from Austria, the Movimiento contra la Intolerancia (MCI) from Spain, jugendschutz.net from Germany, the Ligue Internationale Contre le Racisme et l'Antisémitisme (Licra) from France, the Inter-Federal Centre For Equal Opportunities and Opposition to Racism from Belgium (now called Unia), and the Magenta Foundation from the Netherlands (MDI); who provided most of the data this report is based upon.

Legal Disclaimer

This publication has been produced with the financial support of the Rights, Equality and Citizenship (REC) Programme of the European Union. The contents of this publication are the sole responsibility of the International Network Against Cyber Hate and can in no way be taken to reflect the views of the European Commission.

Table of Content

I. Introduction

The internet is arguably the most important innovation and discovery of the 20th century. It revolutionised the media, bureaucracy, the financial markets, warfare, politics, the way information travels globally, and people's social lives. We are not just passive consumers of information anymore, but participants in the dissemination of news worldwide and have a previously unheard-of impact on and power over the information we consume. These developments have had profoundly positive effects on our lives. It made life easier, more convenient - and in theory - much more saturated with knowledge.

However, there are downsides to this overhaul that our lives have gone through in the past 20 or so years. A couple of major companies, such as Google, Facebook, Twitter and Microsoft have come to dominate our lives in the online sphere. They have a major influence on what we consume and how we consume it; what we see online and how it is presented to us. This has brought forth several issues. Some of those had already existed in the age dominated by the printing press, some are new. Information bubbles and the consumption of opinions that we already agree with is nothing new. Media sociology already described these phenomena in the second half of the 20th century. However, there has never been an era where this basic human trait would have been catered to so precisely and exclusively. Social media and other internet companies do everything in their power to make us consume more information on their websites and other platforms, since that is how they collect data on people and present advertisements that are more and more fine-tuned to us personally due to the aforementioned data collection. Hence, they want to cater to our needs, so we will consult their platforms and use their services on a daily basis. To achieve this, they use algorithms that increasingly show content personally tailored to our likes and dislikes, creating massive echo-chambers that are becoming more and more impenetrable by outside opinions and information that might challenge our views or make us think outside the box.

These are issues in and by themselves, but it needs to be realised that the problem is much more diverse, deeply rooted and serious than just people being one-sidedly informed about politics or societal issues. The internet has become a second world that is just as real as the world away from the keyboard. It created a global village that is obviously fragmented, but also tremendously interconnected. And since this fragmented and interconnected global village is being created by people, it has become a mirror image of "real life" with all its ugliness, violence and discrimination. This means that racism, hatred, antisemitism, homophobia, Islamophobia, xenophobia, bullying and such are prevalent and rampant on the web. These phenomena are collectively called cyber hate, a fundamentally 21st century issue that needs to be addressed and combatted at every possible turn. Especially because - due to the aforementioned echo chamber effect - people can create bubbles online that are filled with vile hatred towards "the other", whilst being completely unchallenged. This can lead to indoctrination and radicalisation on a previously unseen scale. This is particularly worrying because children and teenagers are amongst the main targets of extremist groups and far-right/far-left parties or pages; trying to grab and shape young minds to further their extremist political agenda.

Fighting online hate speech, however, is no simple task. There are several issues that activists, NGOs and governments have to face when entering this field. Answering questions such as “what do we need to fight?”, “how we want to combat it?”, “how can we enforce laws online?” and “how can we police a practically infinite virtual public sphere AND still respect and maintain human rights such as the freedom of speech?” are all questions that need answers.

The International Network Against Cyber Hate (INACH) has been involved in this field for more than fifteen years, and these questions are still having to be answered by our colleagues at all our member organisations on a daily basis. However, we, at INACH, fundamentally believe that one cannot put a single human right - free speech - on a pedestal above all other rights, just because people interact within the confines of the online public sphere where the freedom of expression is paramount, and ignore all other human rights that are otherwise taken for granted in most civilised nations in the offline world. Human dignity, to be free from degradation, to be free from violence and not to be discriminated against are all basic human rights enshrined in several national laws, constitutions worldwide and many international treaties. Yet, online we tend to ignore them and preach the supremacy of free speech above all else. We believe that fighting cyber hate and trying to get hateful content off the internet is not a limitation of human rights, especially not in a general sense. We actually bring the online in line with human rights (as our motto states), by working towards a cyber space that does not discriminate against members of the most vulnerable communities. On the one hand, hundreds of millions of people are being bullied and discriminated against online daily, turning their experience into a nightmare. On the other hand, millions and millions of people are being slowly indoctrinated by racist and revolting propaganda that is becoming extremely hard to challenge and counter due to the echo-chamber effect.

Therefore, INACH has been monitoring cyber hate for years, trying to get instances of illegal and/or harmful hate speech removed from online spaces, especially social media. Our network has also been meticulously collecting data on cyber hate since the middle of 2016 in the framework of Project Research - Report - Remove: Countering Cyber Hate Phenomena. This Annual Report is the product of this project and will present the drivers, trends and tools that have underpinned online hate speech in the past twelve months. It will also present data provided by our members and project partners on online hate speech, such as the prevalence of hate types, the dominance of different social media platforms and the removal rates of internet companies. It will then present trends within the data in order to provide an extensive and in-depth picture of the phenomenon that is cyber hate. Lastly, recommendations will be put forth.

II. Methodology and Issues Faced

The data collection for this report took place between May 2016 and May 2017 from all project partners residing in multiple EU countries (Austria, Belgium, France, Germany, the Netherlands and Spain). INACH secretariat collected data from these partners on a monthly basis using both quantitative and qualitative methods.

The qualitative data was collected through a Microsoft Word document that asked the following questions from the partners:

- Please provide a short paragraph about emerging or new drivers (e.g. refugee crisis, Daesh terrorism, etc.) of cyber hate in your country.

- Please provide a short paragraph about emerging or new trends (e.g. new target group, growing role of a certain online platform, growing hate against a certain community, etc.) within cyber hate in your country.

- Please provide a short paragraph about emerging or new tools (e.g. memes, conspiracy theories, fake news stories, etc.) used by people to spread cyber hate in your country.

- Please tell us about conferences you organised, campaigns you launched, reports or papers you published on cyber hate; and counter-narratives or counter-speech you use to combat the phenomenon.

As one can see, these questions cover most of the first half of this report that discusses emerging drivers, tools used by extremists to spread hate online and societal trends that can be observed within the field of cyber hate. It also provides the backbone of the chapter about our activities at the end of this report that sums up our partners' fight against online hate speech.

The quantitative side of the data collection was done through a Microsoft Excel table that collected numerical data from our partners based on the cyber hate cases they closed during the previous calendar month.

Through this table, INACH secretariat collected data on hate types, i.e. how many cases a certain partner had in that month that falls under some pre-set umbrella terms. Due to our project's methodology in general, these umbrella terms were the following:

- Racism
- Xenophobia
- Anti-Roma hate
- Anti-Muslim hate
- Anti-religious hate (anything but Islamophobia)

- Hate against non-religious people
- Anti-Arab racism
- Antisemitism
- Anti-refugee hate

Since some cases are not clear-cut and can fall under multiple hate types, INACH and its partners decided to include such cases within all hate types that they fit under and then count them as two or three cases (or as many cases as hate types they fit within). For instance, if a case was antisemitic and homophobic, it was included in the data set under both antisemitism and homophobia and then counted as two cases in the combined number of cases for that month.

The second category that INACH collected data on was the number of cases on different online platforms. We recorded cases on Web 1.0 and Web 2.0 platforms both separately and together. These platforms are the following:

Web 1.0	**Web 2.0**
Websites (comments on websites included)	Facebook
Forums	Twitter
Blogs	YouTube
	Google+
	Instagram
	Vimeo
	Dailymotion
	Tumblr
	Pinterest
	Snapchat
	Telegram
	VK.com
	Other

The third category was the legality of cases that our partners handled. Online hate speech is a very contested phenomenon for obvious reasons. Hence, all nation states and supranational bodies handle and regulate instances of hate speech differently. Also, NGOs, such as our partners, often find themselves in situations where there is online content that is highly offensive, discriminatory or hateful yet it does not violate the laws of the country they reside

in. That is why including this category was important, to give a picture of unsanctioned cyber hate in order to highlight loopholes in the legislature. Therefore, in this category, INACH collected the number of cases that were deemed illegal by the national law of given country and cases that were not deemed illegal.

The fourth category was the actions that our partners took against instances of cyber hate. This included the following subcategories:

- Sent to police
- Sent to prosecutor's office
- Sent to other state authority
- Request for removal
- No actions taken

Just like with the hate types, some cases fell under multiple categories, thus they were included in all categories they fit into and then counted twice or thrice depending on how many categories they were included in.

The fifth and final category of INACH's data collection were the number of removals and non-removals on all platforms that have been mentioned above.

Within all these categories, INACH also produced percentages and thus ratios. Hence, we know the ratios of different hate types, the ratios of the prevalence of different platforms and the ratios of the removal rates on different platforms.

Although everything might seem clear-cut and straightforward based on this chapter so far, the issues that INACH has faced during the data collection period must be mentioned here. The most pivotal problem that INACH had to face and solve is the extreme volatility in the numbers collected. Our project partners have differing focuses within the field and their capabilities and funding differ vastly from one another. Hence, INACH secretariat received higher number of cases from some partners and lower numbers from others. Also, the collected data is influenced by the focus of the different partners. Some are more focused on antisemitism, others are more focused on anti-Muslim hate, etc. Furthermore, the data INACH collected is not anchored to any outside phenomenon and it is not controlled but perfectly random. We did not control the incoming numbers in any way and we did not tie the numbers to any outside factor, such as demographics. Hence, weighting the data was impossible, because it would have been too arbitrary and the weights would have had to be measured for all partners, for all months and for all categories, causing the data analysis to be too chaotic and almost based on happenstance. Therefore, weighting the data was ruled out by our analysts and INACH decided to use moving averages to smooth out the figures and be able to unearth trends from the volatile data pool.

It was decided to use a 4 point interval for calculating the moving averages. The reason for this was twofold, one of our partners – 7ugendschutz.net in Germany – ran an extra

monitoring project unconnected to this project and, therefore, their numbers were outlandishly high for two months during the summer of 2016. These outlying data points had to be smoothed out. Also, our analysts found that a 4 point interval is the sweet spot between the data being extremely volatile and therefore hard to analyse and the data being far too smoothed out artificially and thus representing reality less than ideally.

Due to the aforementioned issues and solutions, the trends discussed in this report are based in the moving averages of the ratios of the different hate type categories and removal or non-removal rates. This way, INACH believes, that our data give a fairly good overview of cyber hate and trends within the phenomena for Europe, especially Western Europe. Sadly, INACH does not have Eastern European members participating in this project, so our data is a bit lopsided as far as the continent goes. Yet, we firmly believe that conclusions can be drawn on the phenomenon in general on a European level, on trends within the phenomenon and on the practices of social media companies when it comes to content removal from their platforms.

III. Drivers, Trends and Tools

As it was mentioned earlier, throughout the year, we collected monthly inputs from our members, namely data illustrated by drivers, trends and tools. The drivers, being real world events, stood behind the emergence of certain trends in online hate speech, which were, in turn, spread, emphasized and enabled by particular tools. The main examples that came forth in each category will now be summarized. As drivers and trends are directly connected, representing the cause and the effect, they will be in the same subchapter, tools being in a separate one.

1. Drivers and Trends

This year was filled with events that shaped and transformed the world as we know it, leading to the emergence of many different trends. The main trend that has been observed is undoubtedly the rise of extremism in the western world. For extremism to rise, it uses scapegoats in order to rally people to the cause. That is where racial discrimination and hate speech against minorities such as Muslims, Jews or refugees emanates. Actually, as it will be observed later on when looking at the analysis of the data, the four types of hate that were at the top of the list this year were indeed racism, antisemitism, anti-Muslim hate and anti-refugee hate. Events such as the transformative elections around the world, notably of Donald Trump in the US, and record number of votes for Wilders in the Netherlands and for Le Pen in France embody this rise. It has been building up for a while, as the multitude of terrorist attacks lead to an atmosphere of fear and paranoia, and as the refugee crisis has led to an even bigger gap between the “us” and “them”. The link between offline events such as those drivers, and the trends online is undeniable. Hence, it is necessary to look at some of the many examples of those drivers to understand the trends that came forth this year. Below are listed a few of those drivers.

Doubtlessly, the terrorist attacks that took place in Europe and in the United States had profound effects. Many of our partners denoted the direct effect that these attacks had on the increase of certain types of online hate speech. Regarding the Nice attack, Germany indicated that right-wing parties instrumentalised the attack to convey their anti-migrant message. The same thing happened in France, with the Front National (the extreme right party) that also exploited the tragic event. The attack in Orlando led to increased online hate as well. On one hand, anti-Muslim hate was fuelled by the attack, which was notably observed in Spain and Germany, where the idea that Islam was a homophobic religion that could not coexist in a diverse democracy arose leading to the escalation of hate speech against Muslims. On the other hand, the attack also fuelled homophobic hate speech, which was particularly notable in Spain and in Belgium.

Of course, non-terrorist attacks also took place throughout the year. Those that were in some way or another linked to Muslims, refugees or members of other minority communities, were also used to feed hate speech concerning such groups. In fact, extremists go as far as blaming all refugees and Muslims for those types of attacks. This ties to the whole scapegoat idea mentioned above, which leads to unfair, false and bias generalizations such as, that, for instance, all Muslims are terrorists. Some of the many examples of this were the couple of deadly attacks that took place in July, in Germany. A teenage refugee attacked people with an axe on a train in Würzburg, a German-Iranian teenager killed and harmed several people in a shooting in Munich, a Syrian refugee killed a woman with a machete and injured several others in Reutlingen, and another Syrian refugee committed a suicide bombing at the entrance of a music festival in Ansbach. On social media, these attacks were interpreted as the arrival of Islamic terrorism in Germany.

Likewise, politics had an undeniably major role in explaining the data collected. For instance, one of the main drivers behind the observed trends this year was the context of the presidential elections in Austria, the Netherlands, France and Germany. In May, as the final round of Austria's presidential election took place, hateful comments increasingly appeared in online discussions about the election and politics. In addition, death threats against the newly elected president was posted on the Freiheitliche Partei Österreichs (FPÖ) leader's Facebook page. In France, there was also an increase in extremist and radical political statements. The rise of far-right ideas in France was firmly visible, whilst racist and extremist statements from politicians were trivialised and normalised, in turn leading to an increase in anti-refugee and anti-Muslim hate. Brexit was another driver which had an international impact, namely on Spanish politics as it was used by political parties during the general elections to instigate heated debates, once more leading to online hate. Furthermore, Donald Trump's election campaign supported by the so called "Alt-Right" was part of the reason for this increase in hate speech in Europe. For instance, "Pepe the frog", one of the most popular symbols of the Alt-Right, was used by members of the anti-immigrant Identitarian Movement and German right-wing extremists as part of their social media strategy. Moreover, the attempted coup in Turkey strengthened already existing anti-Muslim, anti-Turkish and generally racist resentments, especially in Belgium, Austria and the Netherlands. Another driver was the burkini ban in France, which reinforced demands in Austria to also prohibit

religious, mainly Muslim, clothing in public spaces. Besides, in Germany, in September, protests against refugees in Bautzen took place. However, most of the protesters were not “concerned citizens” but neo-Nazis. After violent confrontations between refugees and protesters, far-right activists tried to instrumentalise the events and instigate hatred against refugees. Other than that, there are also drivers that reappear every year at the same time. In the Netherlands, in December, the Zwarte Piet (Black Pete) debate heated up again, as it had for the last 5 years. Online and offline demonstrations were held, both by those against Zwarte Piet because of black face and all the racism that comes with it; and by those pro-Zwarte Piet, as it is - according to them - a innocent holiday tradition. Incitement to hatred and violence took place and MDI received several complaints from both sides around that time. Still, in the Netherlands, incidents surrounding a new political party called “Denk”, which focuses on fighting against discrimination and racism, appeared in November. The two Turkish founders and especially Sylvana Simons, one of the leaders of the party, became victims of hatred and discrimination. A video was made and pictures were posted online in which black people hanging from a tree were photoshopped with a picture of Sylvana’s head on them. The video was accompanied with an upbeat carnival song about the fact stating that Sylvana needed to leave Holland. It is important to note, however, that the Turkish founders were later involved in the spreading of anti-Dutch and antisemitic remarks accompanying their pro-Erdogan statements.

One last example of a driver that is separate from politics was the European Championship of the summer of 2016. In Germany, it was noticed that this Championship led to many discussions within the far-right movement, especially regarding the multiculturalism of the team, as the concept of a "national" team was – in accordance with the general discussions about the decline of the "real" German people – considered a failure.

2. Tools

After the above overview of which trends emerged due to which drivers, it is essential to look at which tools were used in order to facilitate the surfacing of aforementioned trends. These tools are either online or offline. Concerning online ones, the close to unlimited freedom of creation on the internet is important to keep in mind in this context. Every day new tools arise, whether they consist of images, videos, articles, and whether they are fake or not does not matter. Their goal is to incite hatred in people and tools, such as fake news have devastating consequences. As the speed of sharing is unprecedented, the harm it can do is unmatched.

Fake information was one of the most widely used tools. For instance, in Germany, in the summer, there was a story of a knife attack at the train station in Grafing (near Munich), which spread via social media immediately. Although it was clear shortly after that the attacker had no terrorist background, innumerable hoaxes were emerging on internet. Even when the investigative authority declared that the perpetrator was an autochthonous German who was known to the police as mentally confused, right-wing extremists carried on propagating that this was a misinformation campaign. Furthermore, Marco Delgardo, a

popular former music producer, claimed on his blog that the perpetrator's name was Rafik Youssef. Despite the obvious falsehood of the statement, the post was shared several thousand times and taken up by "alternative media" outlets, like KOPP-Verlag and COMPACT-Magazin. Similarly, conspiracy theories also surfaced. Here are a couple of examples. In France, during the summer, a conspiracy theory related to the European Championship of 2016 emerged. Benzema and other football players of Muslims origin had supposedly been rejected from the French football team because the Jewish lobby refused to let the Muslims be part of it. Another conspiracy theory related to Brexit, being that Jews were supposedly behind Brexit in order to destroy the White European Civilization. Fake photos were also used to convey hate. An example, out of the many, was the fake photos which were posted online in France of a man presumed to be a refugee wearing a T-shirt with "fear for your wife" written on it. The photo was shared by the extremists and was retweeted 200 times in one day. It turned out that the photo had been taken in Australia in 2013, so it had nothing to do with refugees. Nevertheless, there were also real photos used for hateful purposes. For example, in March, after the attack in London, a picture circulated on far-right social media sites (especially on Austrian and German ones), which showed a woman wearing a hijab looking at her phone on Westminster Bridge as people gathered around an injured person nearby. The picture was posted with texts such as "Muslim woman pays no mind to the terror attack, casually walks by a dying man while checking her phone" and incited a vast number of appalling hate comments against the woman and Muslims in general.

Besides fake news, other tools were also used, especially on social media. In France, for instance, #GreatReplacement appeared. The theory of the "great replacement" was developed by Renaud Camus and stands for "the colonization of France by Muslim immigrants from the Middle East and North Africa, which threatened to "metamorphose" the country and its culture permanently". According to him, this theory was proven by the "proportion of Muslim children" in classrooms. Once, for example, he tweeted "#StartOfSchool: the most obvious proof of the ongoing #GreatReplacement". Facebook live, a feature which lets people, public figures and pages share live video streams with their followers and friends on Facebook, was another tool used to promote hate speech. This was the case for the German activist group called "Matefaschisten", who streamed a live video via Facebook on December 15th. Their statements oscillated between fooling around and neo-Nazi propaganda, as one participant said, "Bombs on Israel!" for instance. In this case the activists only removed the content after the broadcast was finished.

On the internet, not only social media is used to fuel hatred. For instance, Amazon.com emerged as a tool to accomplish this aim. LICRA found that people were selling a T-shirt of Pikachu dressed as Hitler. The sale was removed 6 days after LICRA's request. LICRA also made a referral against the far-right and antisemitic rapper, Amelek, who is a member of this new trend of "identitarian rap", promoted by well-known antisemitic figures, such as Dieudonné or Alain Soral. Memes were other examples of hate spreading tools. In Germany, the so-called "orc-postings" memes emerged. These memes, derived from the Lord of the Rings franchise, showed refugees as orcs. The idea of "orc-postings" originated from the American "Alt-right" movement but quickly spread through the country in May. The Press

was another tool to spread hate. In May, the 9th edition of the Islamic State's magazine "Rumiyah" was published in German. It contained anti-Christian and anti-Shia content as well as a step-by-step guide for truck attacks. Chain emails also re-emerged, notably in Belgium, in June, to promulgate online hate. UNIA noted that chain mails were relatively frequently used as a tool in the past, but had somehow disappeared, until 2016-2017, when they reappeared in force used by extremists to spread their propaganda. Hashtags were another tool used to promulgate hatred.

Regarding offline tools, they are prevalent and harmful as well. Campaigning was one more tool used to spread cyber hate. In a symbolic campaign in Germany, in May, the identitarian movements of Germany and several other countries blocked the way of rescue ships in the Mediterranean Sea. They claimed that NGOs rescuing refugees from drowning were in fact human traffickers and that less refugees would come to Europe if more would drown in the sea. The campaign was designed to resemble campaigns of environmental activists like Sea Shepherds or Greenpeace.

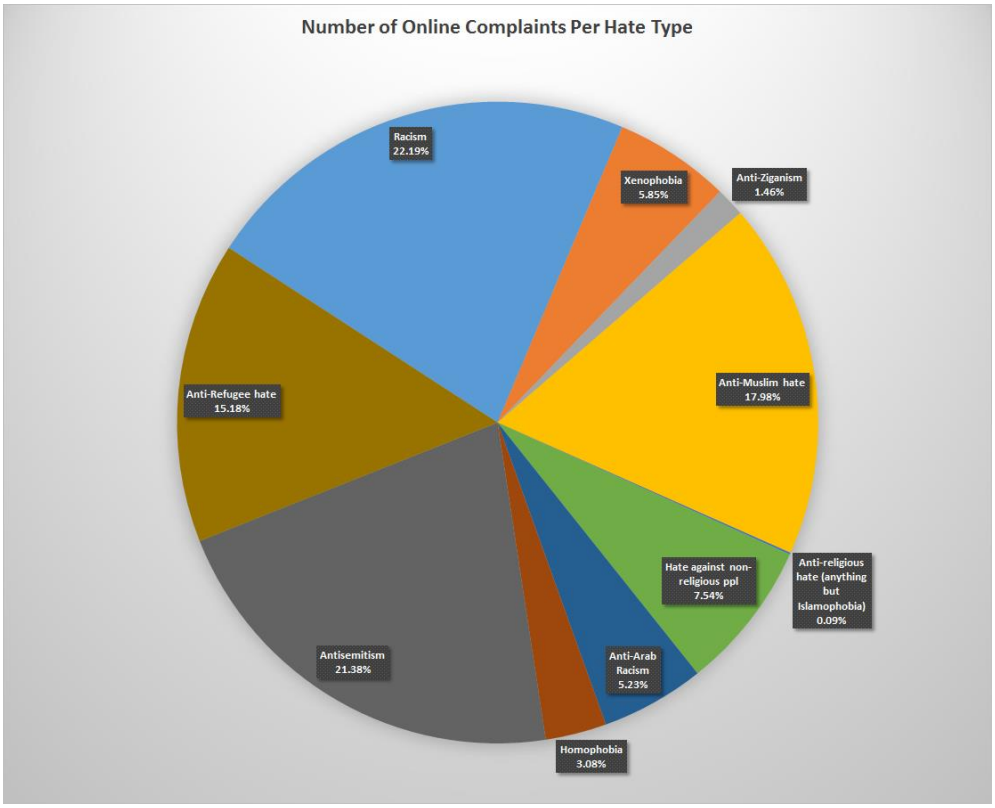
The existence of all these tools enable us to have an understanding as to what facilitates the spreading of the hate which is embodied in trends that emerge through drivers.

IV. Data

1. Introduction of the Collected Data

In this chapter, a snapshot of the data will be given to familiarise the reader with the situation in Europe as far as cyber hate goes. All categories introduced in the methodology chapter will be introduced and discussed here. However, the reader must keep the issues discussed in the methodology chapter in mind, since some of the conclusions that will be drawn in this chapter based on the data collected in the past year are heavily influenced by those issues; such as the manpower of our partners and the type of cyber hate they are focusing on. Still, this chapter will provide an extensive and in-depth overview of the phenomenon.

A) Hate Types



Based on our data collection, four hate types can be seen as predominantly prevalent in Europe, especially in Western Europe. These hate types were dominant all throughout the past year, and even though they might have changed places from one month or quarter to another, their place in the top four was never really in question. These hate types are the following: Racism (22.19%), antisemitism (21.38%), anti-Muslim hate (i.e. Islamophobia) (17.98%) and anti-refugee hate (15.18%). Some of these hate types are showing downward trends, but they are and have been far above all other monitored hate types throughout the year. We will call them the top four hate types.

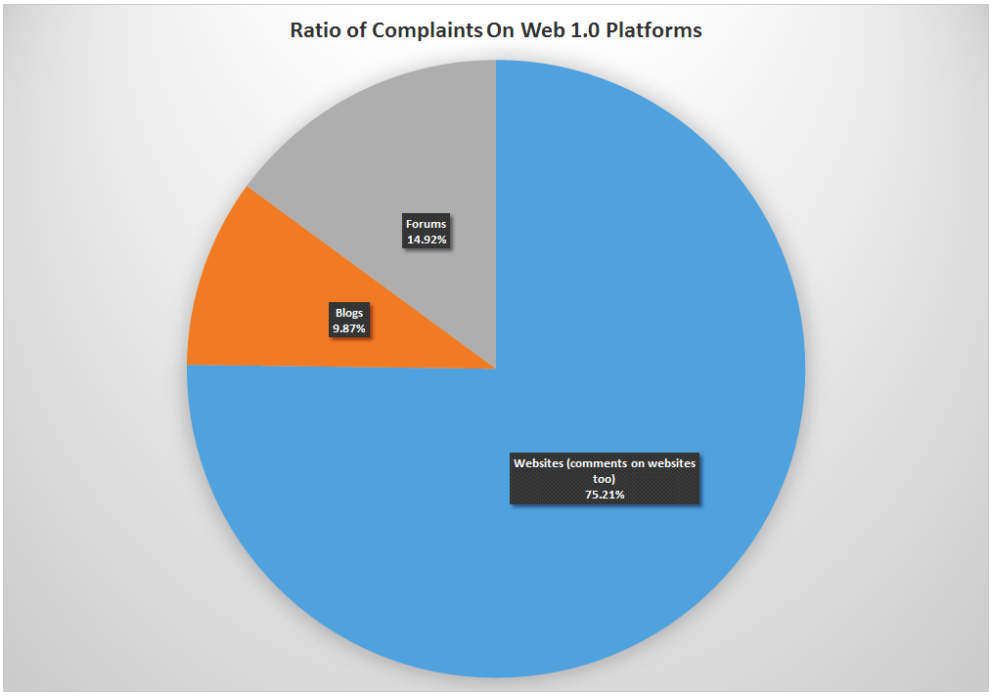
Following the top four hate types, we have the bottom four hate types. These hate types are far below the top ones that make up more than 76 per cent of the data collected by INACH. None of the bottom four hate types reach 10 per cent and only two of them go above 5 per

cent, just. These hate types are the following: Xenophobia (5.85%), anti-Arab racism (5.23%), homophobia (3.08%) and anti-Roma hate (i.e. anti-Ziganism) (1.46%).

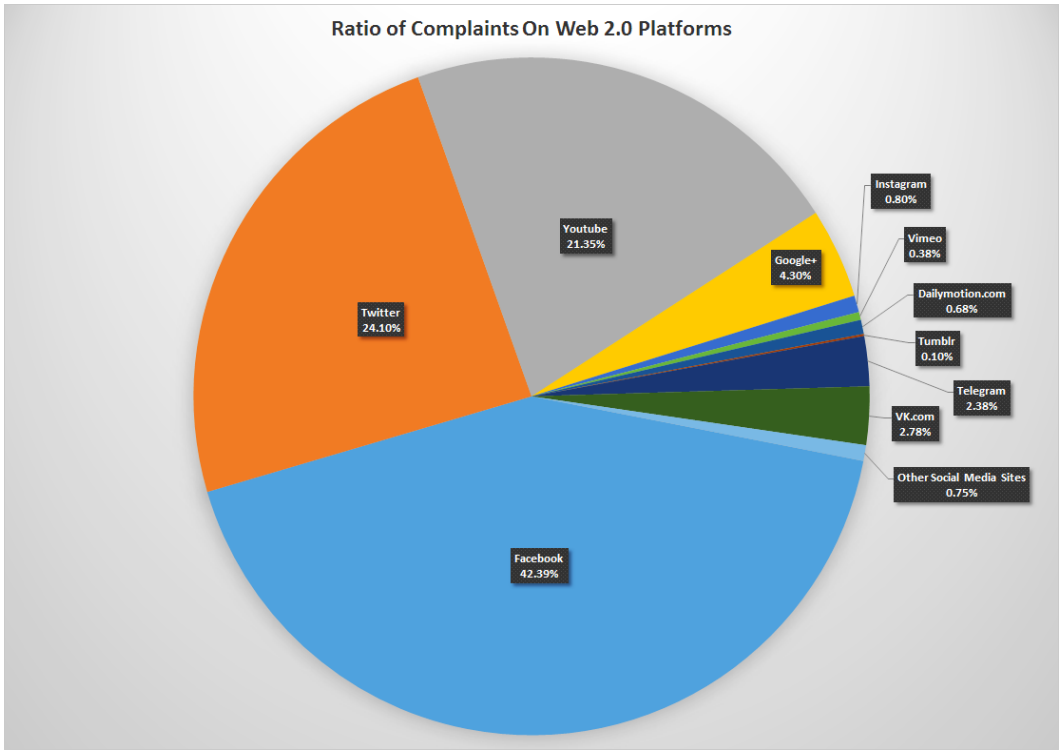
As one can see there are two other categories of hate left (hate against religious people [anything but Islamophobia], and hate against non-religious people). These categories were included per the request of jugendschutz.net, because they collect a lot of cases that fall into these categories. However, due to this peculiarity, more than 98 per cent of the data on these hate types came from Germany. Thus, it was decided to leave these categories out, because they do not represent these phenomena on a European level.

That being said, it is perfectly clear that general racism and antisemitism were the most neuralgic issues in the past year based on INACH's data. Closely followed by Islamophobia and anti-Refugee hate. Anti-refugee hate is a special category that was created out of necessity, even though - as a hate type category - it had been virtually unseen before the so-called refugee crisis that started in 2015. Since then, however, it has clearly been a major issue, although a diminishing one, that will be discussed in detail later in the report.

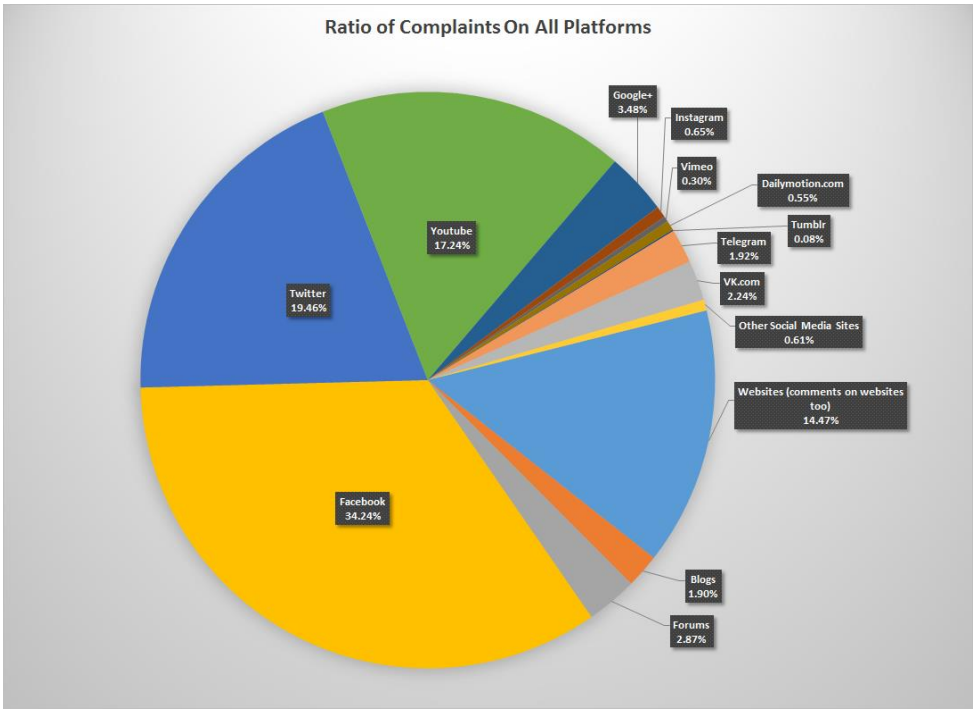
B) Ratio of Complaints per Online Platform



When it comes to platforms that online hate speech spreads on the most, social media is unbeatable. However, Web 1.0 platforms are still in the game. If we take out Web 2.0 platforms from the data pool, one can see that websites are a magnitude above forums and blogs. Three quarters of all complaints registered on Web 1.0 platforms by our partners were registered on websites. They are followed by forums (14.92%) and blogs (9.87%).

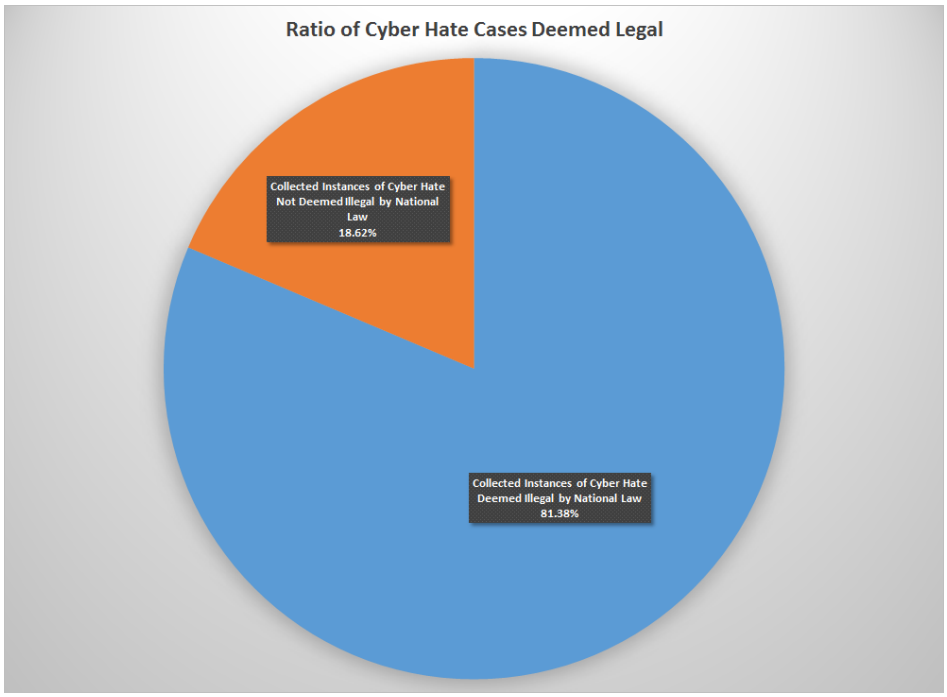


As far as social media platforms go, there is a clear triumvirate that rules the whole market, and therefore gives the biggest surface to cyber hate and extremist propaganda. The three main platforms are Facebook (42.39%), Twitter (24.1%) and YouTube (21.35%). Almost 90 per cent of registered instances of cyber hate came from these platforms if Web 1.0 platforms are taken out of the data set. Facebook's dominance is even more prevalent, since one can see that the ratio of instances of cyber hate registered on it is almost twice as big as the second platform, Twitter. The puissance of the triumvirate is underpinned by the fact that all other social media platforms are dwarfed by them quite literally when it comes to registered cases of cyber hate. The fourth largest number of cases were registered on Google+, which is a meagre 4.3 per cent. Besides the two Russian platforms, Telegram and VK.com that are both above 2 per cent (mainly based on data coming from Germany), no other platform even reaches 1 per cent.



If the two data sets are combined, namely the ratio of cases registered on Web 1.0 and Web 2.0 platforms, it becomes perfectly apparent that social media dominates the online public sphere, and thus most instances of online hate speech are registered on these platforms. Even with the data from Web 1.0 platforms added in, the three social media giants (Facebook, Twitter and YouTube) are responsible for more than 70 per cent of registered cases, insofar as they provided the online space for more than 70 per cent of online hate registered by INACH and its partners. This unparalleled dominance should also come with the highest level of social responsibility that can be possibly taken by multinational tech companies, but this will be further discussed later in this report.

C) Legality of Registered Instances of Cyber Hate

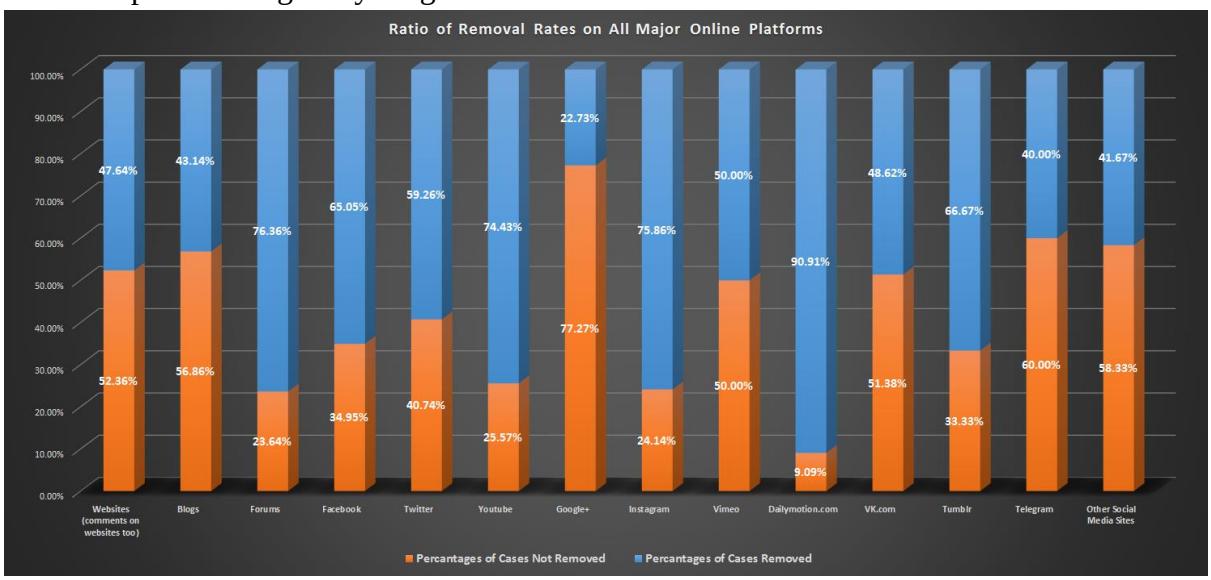


The legality of instances of cyber hate is minor but a quite complicated issue. As it can be seen, the absolute majority of collected cases were deemed illegal by our partners (81.38%). Yet, the fact that

almost 20 per cent of instances of online hate speech collected by the experts at INACH and its partners fall outside of national laws, international directives and EU framework decisions is a bit worrying. It signals gaps in the legislature that ought to be remedied by either EU bodies or the member states themselves. These cases of cyber hate are just as vile, hateful and capable of inciting hatred or radicalise people, yet - due to contextuality or some other loophole in the body of law - they are not penalised in any way, and therefore they are very hard to get removed from the online public sphere. This issue will be addressed later in this report.

D) Removal Rates on All Major Platforms

As far as classical online platforms go, getting hateful content removed from them is extremely hard. This definitely shows in the recorded removal rates. Websites removed 52.36 per cent of questionable content, blogs removed a bit more (56.86%) and forums removed a minuscule 23.64 per cent. The reasons behind this is probably twofold. Minor Web 1.0 platforms are not as prepared or well-funded enough to maintain an army of admins and moderators as social media companies. The second reason, which is also a major issue, is the fact that some of these platforms are specifically brought to life and maintained to give a surface for online hatred. Most of these are hosted on servers in the US and therefore it is almost impossible to get anything removed from them.



However, social media companies with all their money, data and manpower are also far away from perfect. INACH and its partners have to face massive issues due to vague policies and codes of conduct put forth by these companies. They also implement their own rules often highly arbitrarily. Moreover, their stance on different hate types or modes of online hate speech vastly differ from country to country, even though they are supposedly using the same rule book. That is why the numbers we see are not too far from being abysmal. Facebook removed 65.05 per cent of cases, Twitter only 59.26 per cent and YouTube 74.43 per cent. These numbers are fairly low and show the great divide between NGOs that fight for a more inclusive online public sphere and social media companies that try to paint themselves as the knights in shining armour protecting free speech online. However, the fact is that these

companies are money making machines first and foremost and they therefore resent the idea of spending more money to earn less money. And, essentially, that is what NGOs and some governments try to get these companies to do. Higher more people and devote more resources to remove content that - if left online - make them earn money.

INACH, naturally, is not arguing that the removal rate should be a 100 per cent. But INACH has several member organisations that have dozens of experts working for them. These experts are well trained in the recognition of hate speech and international and national hate speech laws. Thus, when they approach these platforms to get something removed, they do that with the knowledge that the content is definitely hate speech AND most likely illegal. Still, there can be differences of opinion, but removal rates should most definitely reflect these facts and therefore they should be somewhere around 90 per cent (at least).

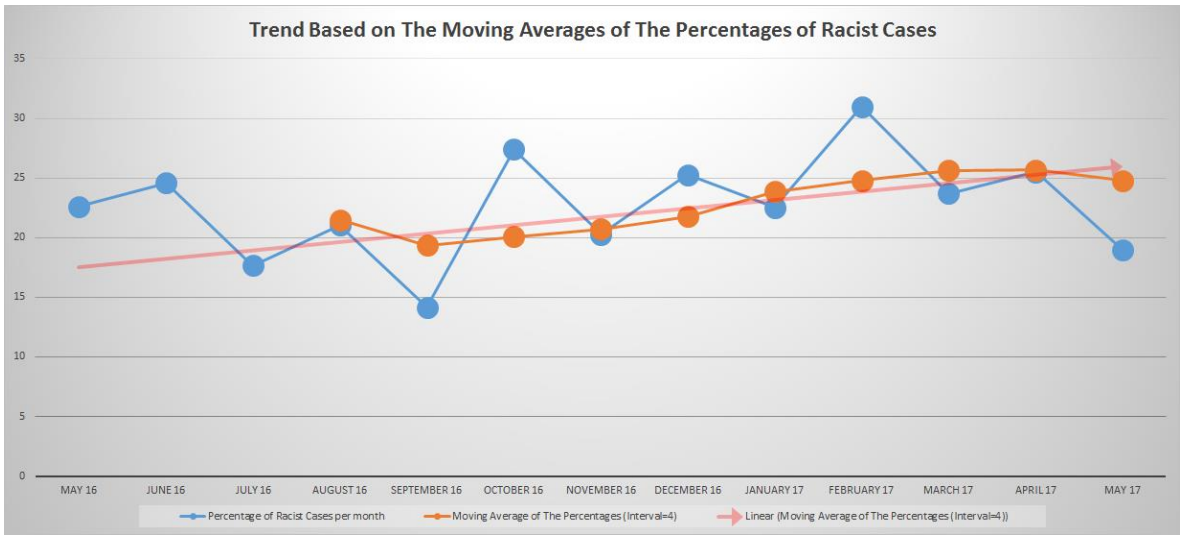
Trends in removal rates, YouTube's removal rate that seems exceptionally high and our recommendations for social media companies will be all discussed in the coming chapters.

2. Emerging Trends in the Data

In this chapter light will be shone on trends that emerged within the field of cyber hate based on INACH's data collection efforts. Trends in hate types, furthermore removal rates on the three major social media platforms will be examined closely to give a general idea about the most singled out targets of hateful online content and the hardships NGOs face while trying to clean up the online public sphere. Some conclusions about the targeted vulnerable communities and the trends within the hate type data will also be drawn.

A) Trends in Hate Types

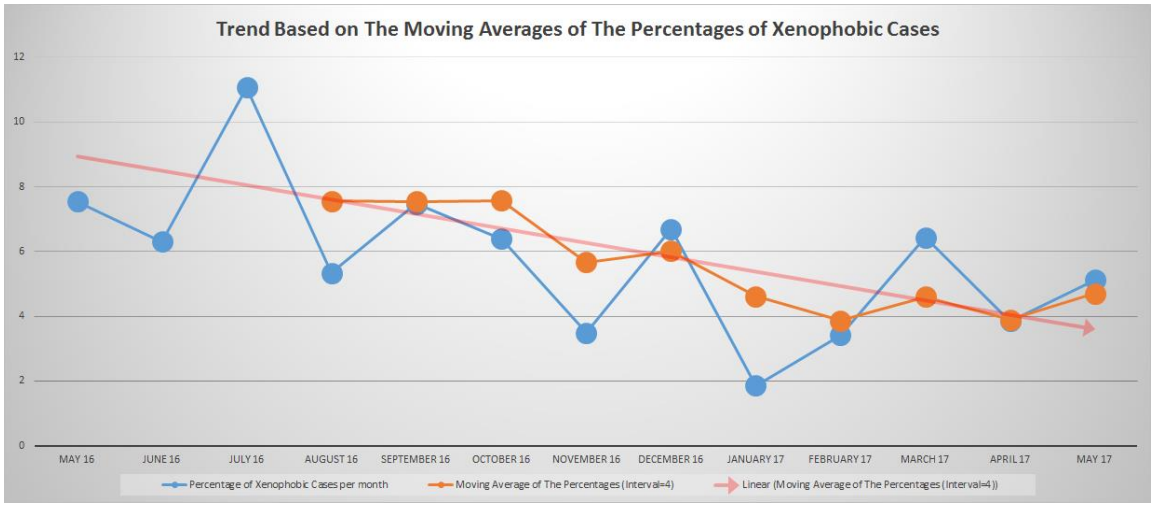
Racism is one of the two most generic hate types within our data set, the other one being xenophobia. However, due to the focus of our project, our methodology excluded several types of hate that otherwise could be classified as racism. Therefore, this data does not include cases that fall into the following categories: Anti-Roma hate, anti-Arab hate,



anti-refugee cases and antisemitic cases. This gives us a narrower picture as far as racism as a category goes but, on the other hand, it gives us a more in-depth and precise picture of the hate types that were examined separately (these will be discuss later in this chapter).

As it has been mentioned, racism is the leading hate type category based on the collected data, as almost the quarter of all cases handled by INACH and its partners fall into this category. This dubious first place is also underpinned by a slow but steady upward trend in racist cases, where the ratio of such instances of online hate speech rose from around 20 per cent on average to 25 per cent on average. Whether this trend will continue during the next year is yet to be seen. However, if one takes a look at the actual ratio of racist cases in the past three months and not just the moving averages, a clear dip can be seen. This dip has not been large enough to turn around the upward trend, but it might signal a change in the near future. Although, since this is one of the two most generic categories and definitely the most prevalent, it is highly unlikely that it would fall further. Hence, INACH is predicting that - on average - the ratio of racist instances of cyber hate will stay around or above 20 per cent and the continuation of the upward trend or the stagnation of the numbers around this level is much more likely than a fall.

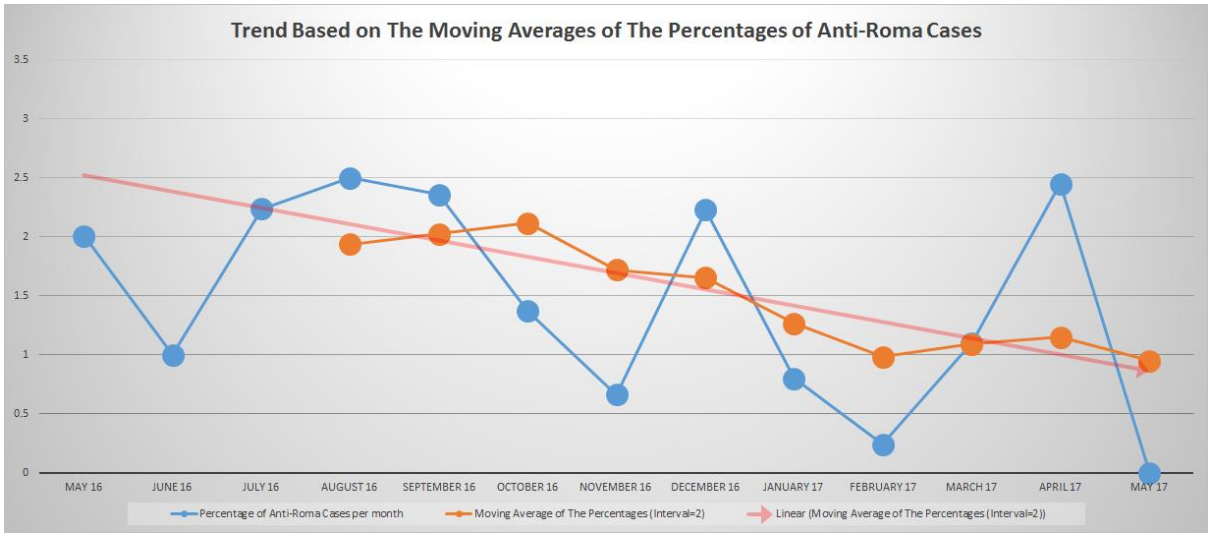
Xenophobic cases fall under the second most generic hate type category that was used during data collection. Unlike racism though, xenophobia is among the four bottom hate types. Its ratio among all hate types hardly ever reached 10 per cent, and on average it stayed firmly under 8 per cent during the data collection period. This fact is also buttressed by a downward trend in xenophobic cases, where their numbers fell - on average - from just under 8 per cent to around 4 per cent, halving their ratio among all other hate types.



INACH's prediction is that xenophobic cases will stay among the four bottom hate types and keep moving up and down within the range they have been fluctuating during the past year. This is probably due to the fact that there are very few cases that clearly fall under the xenophobia umbrella and cannot be categorised as racism or some other hate type. Therefore, its numbers will most likely stay low.

Anti-Roma cases are the first hate type category where issues in our data collections must be discussed. As it has been mentioned in the methodology chapter, no Eastern or Central

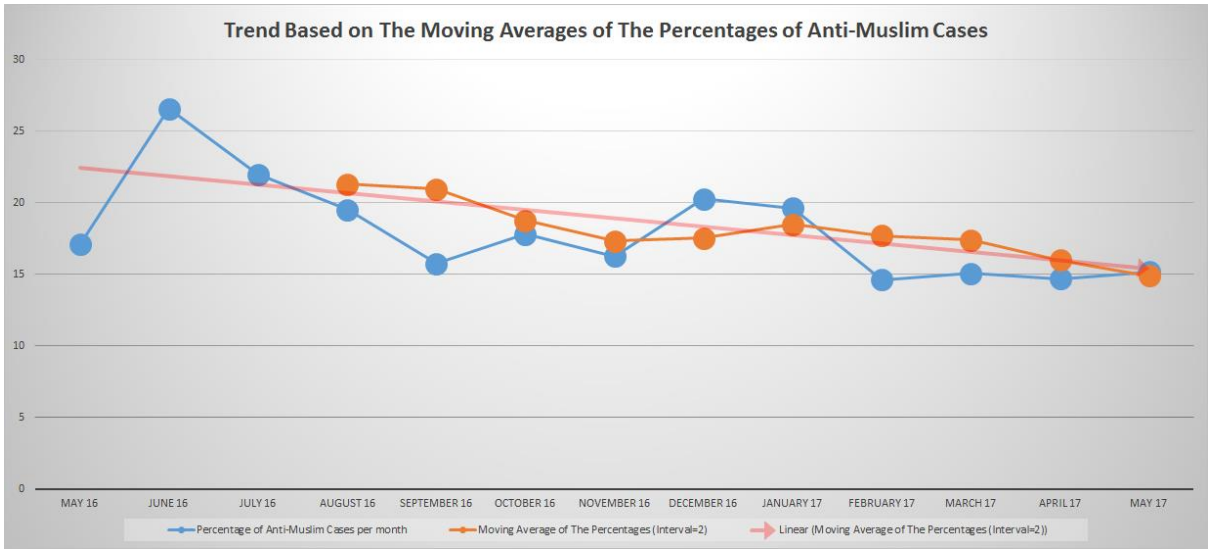
Eastern European members of INACH are participating in the project. Furthermore, during most of the data collection period, INACH was sorely lacking members from the former Eastern Bloc. Moreover, none of the participating partners really focus on hate against the Roma community. Therefore, the number of anti-Roma cases that were collected were quite low.



Now, racism towards the Roma community is not a singularly Eastern European issue, however it is certainly a more prevalent and pressing problem in countries with a high Roma population, such as Slovakia, Hungary, Bulgaria or Romania. INACH has grown since and we welcomed several new members from Eastern Europe. Hopefully our data will reflect that during our next data collection period.

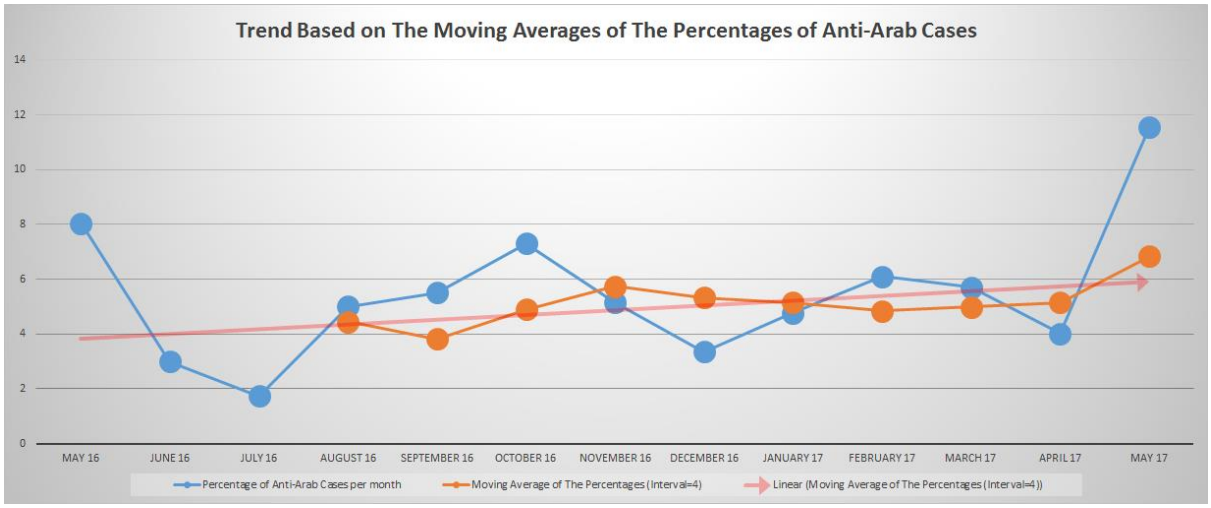
Looking at the line chart above, it is clear that - mainly due to the issues discussed above - anti-Roma cases never really went above 2 per cent on average, fluctuated immensely in an absolute sense and actually fell to zero during the last month of the data collection period. Hence, INACH would not go as far as drawing conclusions based on this data. Hopefully we will be able to collect a larger sample size the next time and give a more well-founded picture of online hate targeting the European Roma community.

Anti-Muslim hate can be found on the completely opposite side of the data spectrum. INACH was able to collect an ample sample size and this hate type category is firmly within the top four hate types. Even though a downward trend can be observed for this hate type, it is also among the steadiest of all hate type categories, on average staying around 20 per cent and then slowly falling to 15 per cent during the last four months.



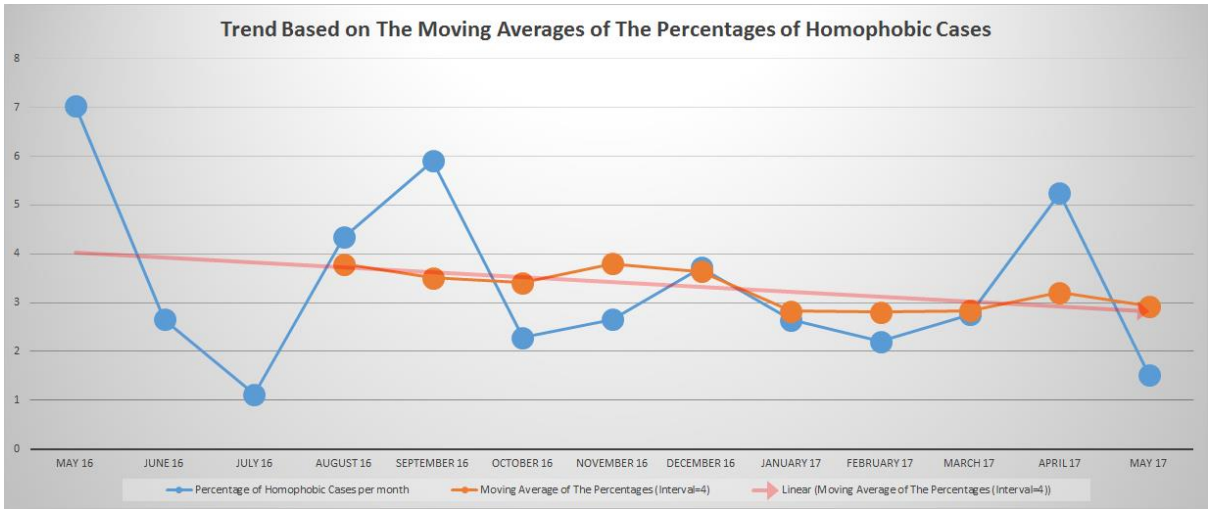
Despite this downward trend, INACH is predicting that the ratio of anti-Muslim cases will most likely not fall any further and will keep on fluctuating between 15 and 20 per cent, staying firmly among the top four hate type categories.

Anti-Arab hate is intimately linked to the previously discussed anti-Muslim hate, however it is also quite separate. It is a collection of cases where people are being attacked solely for being of a Middle Eastern or North African descent. However, it very often overlaps with Islamophobia, since most people and extremists link these two things together.



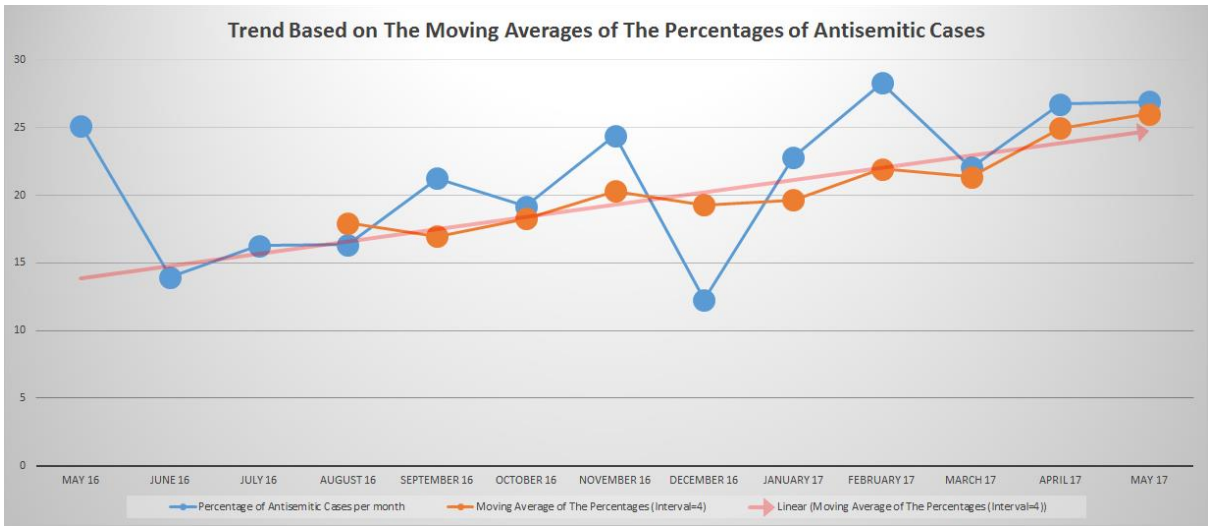
Yet, unlike Islamophobia, anti-Arab hate is in the bottom four and its ratio among all hate types has been around 5 per cent quite steadily with a very slow and small upward trend. This is highly unlikely to change in the near future.

Homophobia is the second hate type after anti-Roma hate that is suffering from a small sample size. None of INACH's project partners focus specifically on homophobia and therefore the collected numbers are fairly low. Still, it can be said that homophobia - as far as our data set goes - is among the bottom four hate types and it stayed very steadily around 3 per cent on average throughout the data collection period.



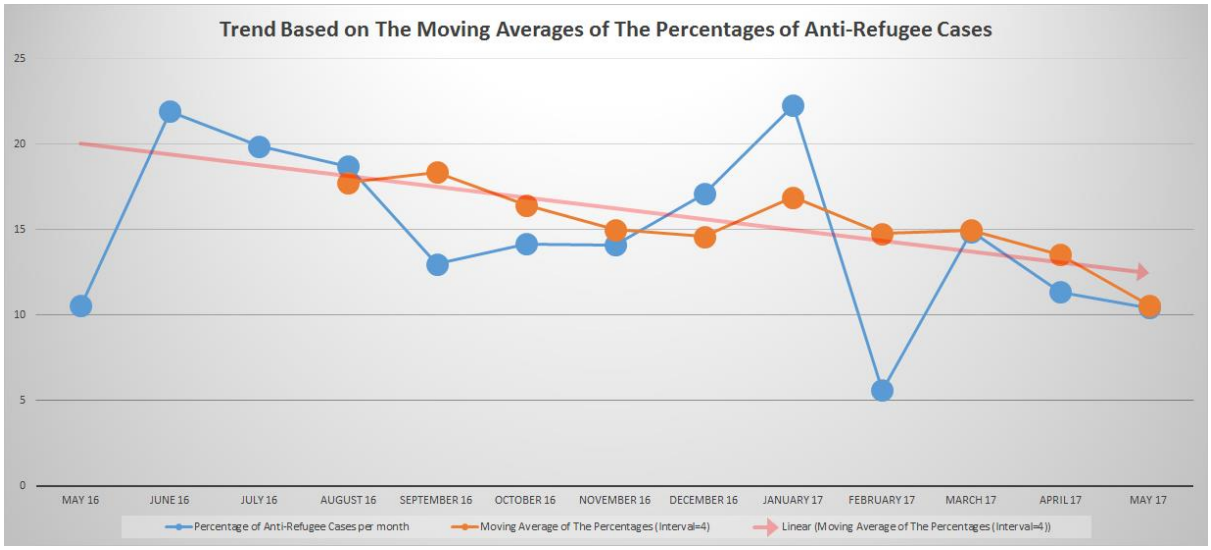
If no additional data will be collected during the next period, it is highly unlikely for this to change, and homophobia will stay at very similar levels to where it is right now.

Antisemitism on the other hand is the second most prevalent hate type within INACH's data set. This is especially worrying because the first one is racism, a very generic and wide hate type, whilst antisemitism is very narrow and specific. However, we have to note here that our French project partner, LICRA, is specifically focused on antisemitism (but not exclusively). Hence, the numbers INACH receives from them are always quite "antisemitism heavy" and therefore the French data skew the sample somewhat, but not to a sufficient extent to make it unusable.

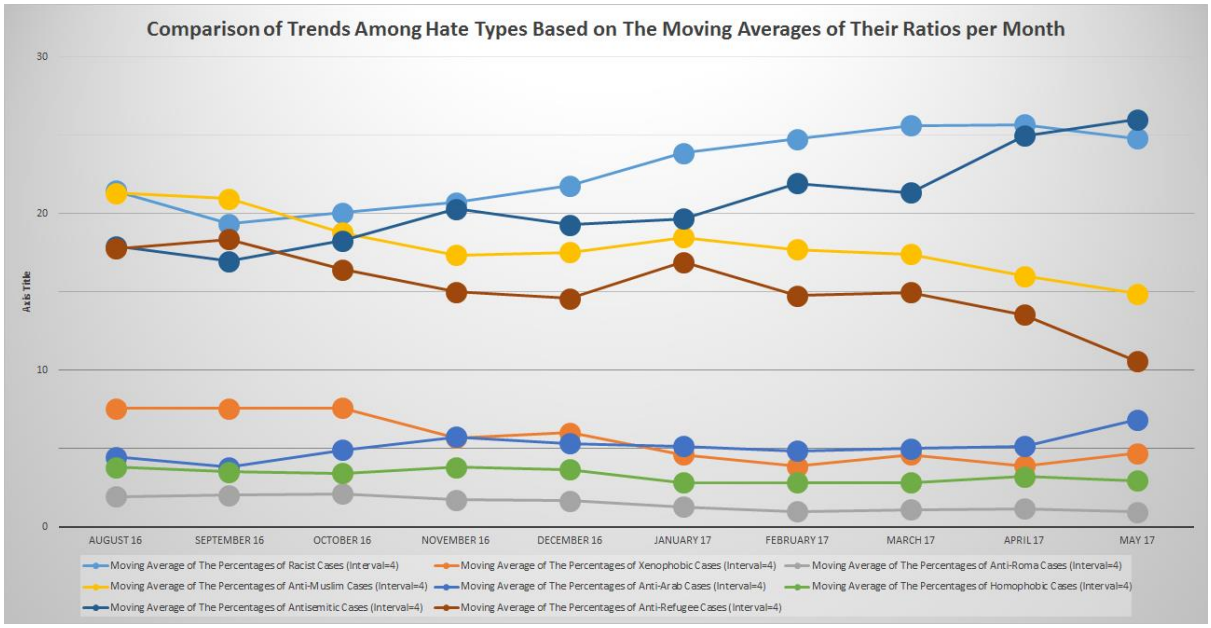


Thus, it can clearly be said that antisemitism is a hate type with a very clear and steep upward trend, that raised this hate type's ratio from around 15 per cent on average to above 25 per cent. This 10 per cent rise is unparalleled by any other hate type in the data set. INACH observed a dip in December 2016, but that dip corrected itself instantaneously and no other sudden change could be observed in the data ever since. This suggests that antisemitism is one of the most prevalent and worrying issues in the phenomenon that is cyber hate, which will almost definitely stay this way in the coming months.

Anti-refugee hate is a new phenomenon within online hate speech. This category contains cases of people being attacked online based solely on the fact that they were refugees irrespective of their religion, sex, ethnic background, etc.



Very little of this type of hate could be observed before 2015, but it had to be included in INACH's data collection because it became one of the most virulent types of hate on the internet by 2016. This was of course due to the refugee crisis that hit the EU during the summer of 2015. This crisis somewhat abated during the past six to eight months, but it still lurks in the background.

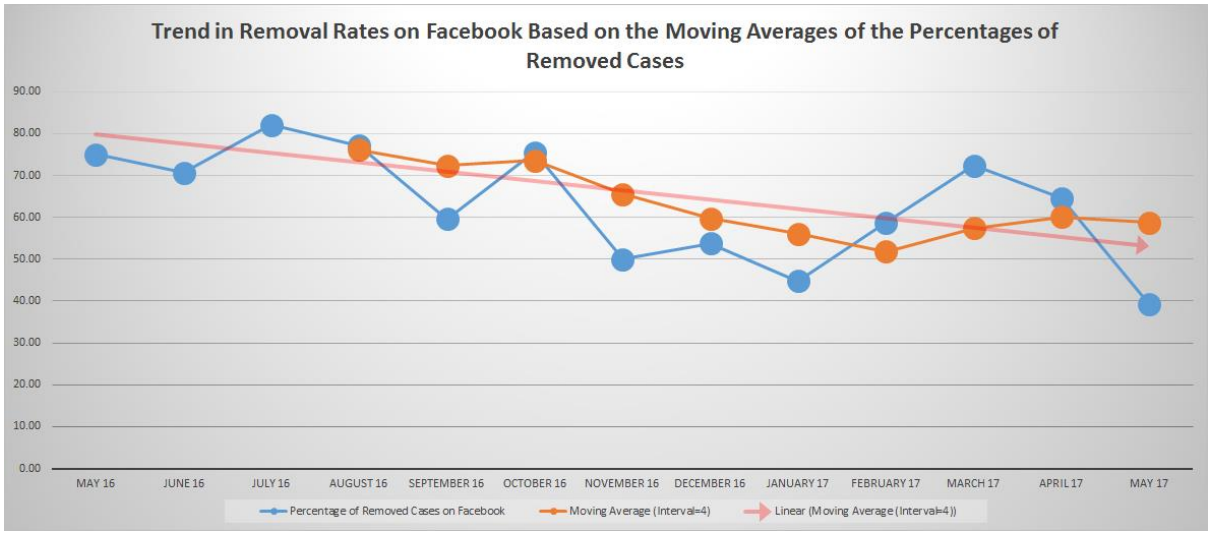


That is why anti-Refugee hate is in the top four hate types according to INACH's data, but it has the lowest numbers among the top four. Especially, because a very steady decline can be observed in this hate type, where it fell from around 18 per cent on average to around 10 per cent, almost halving its ratio among all hate types.

B) Trends in Removal Rates

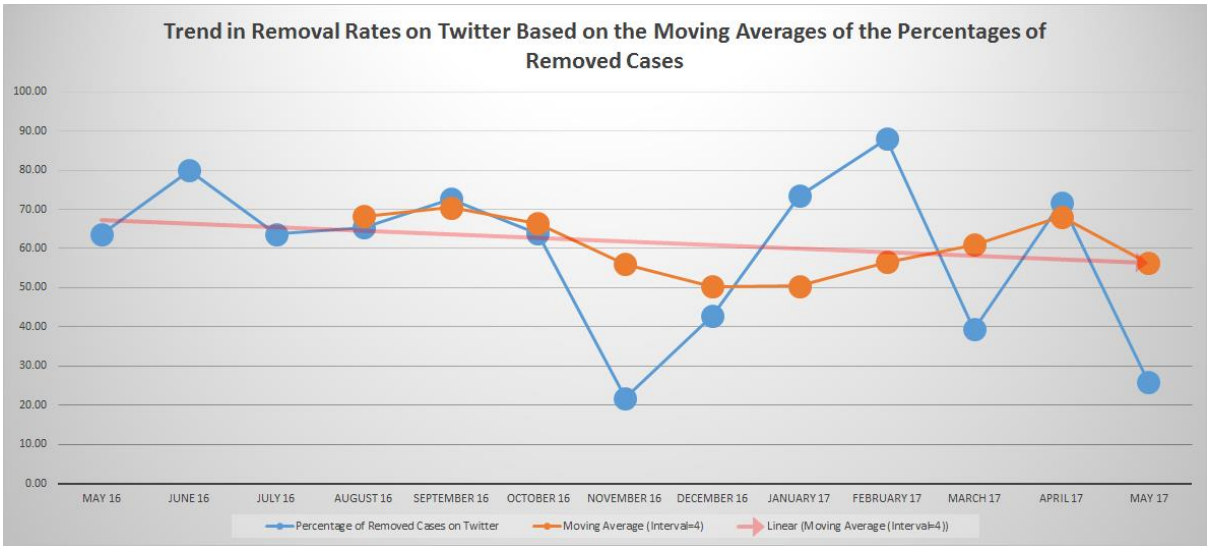
Content removal is among the top goals of INACH and our partners. Cyber hate is corrosive, discriminative and more than capable of radicalising people. Hence, the removal of such content from social media sites is of paramount importance. However, since there is an obvious clash of human rights (between human dignity, freedom from discrimination and freedom of speech) and a clash of interests between NGOs and social media companies. The removal rates on the major social media platforms are far from ideal.

Facebook's removal rates are mostly ok, but they still fluctuate immensely and a slight, but very steady downward trend can be observed on average. As one can see on the chart below

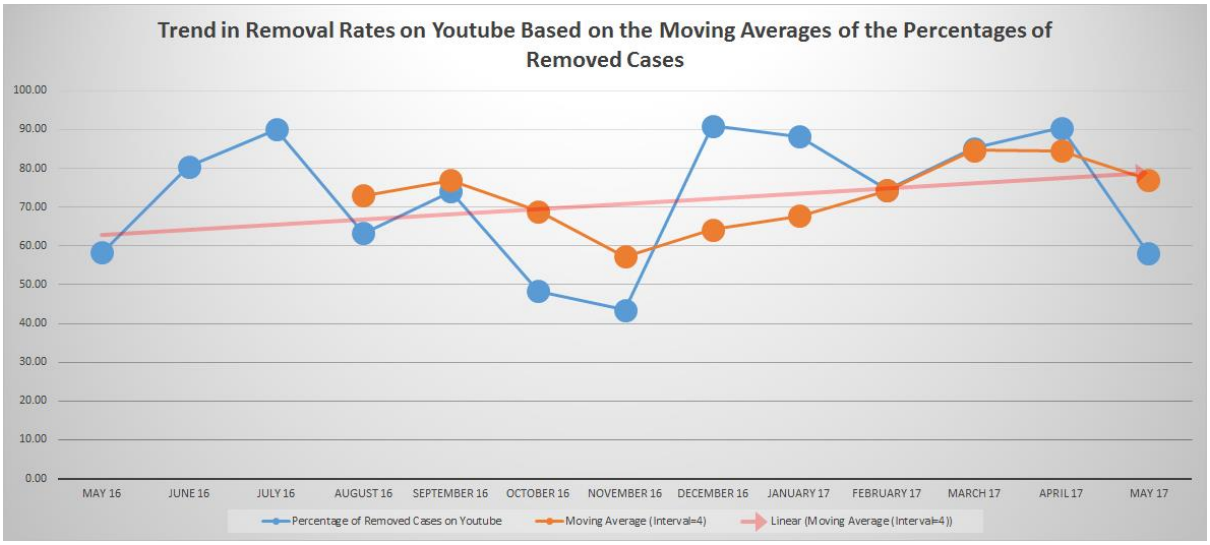


Facebook's highest removal rate was recorded in July 2016 at around 80 per cent. This means that 80 per cent of complaints that our partners sent to the company were removed. However, the lowest removal rate was at below 40 per cent and that was recorded in May 2017. Both this fact and the observable downward trend are bad signs as far as content removal goes.

Twitter is worse than Facebook in removals and they also present a downward trend. Even though their highest removal rate in February 2017 was at almost 90 per cent, their lowest one was at only 20 per cent in November 2016. A ratio much lower than the lowest of Facebook. Moreover, they remove less cases on average and their numbers fluctuate much more.

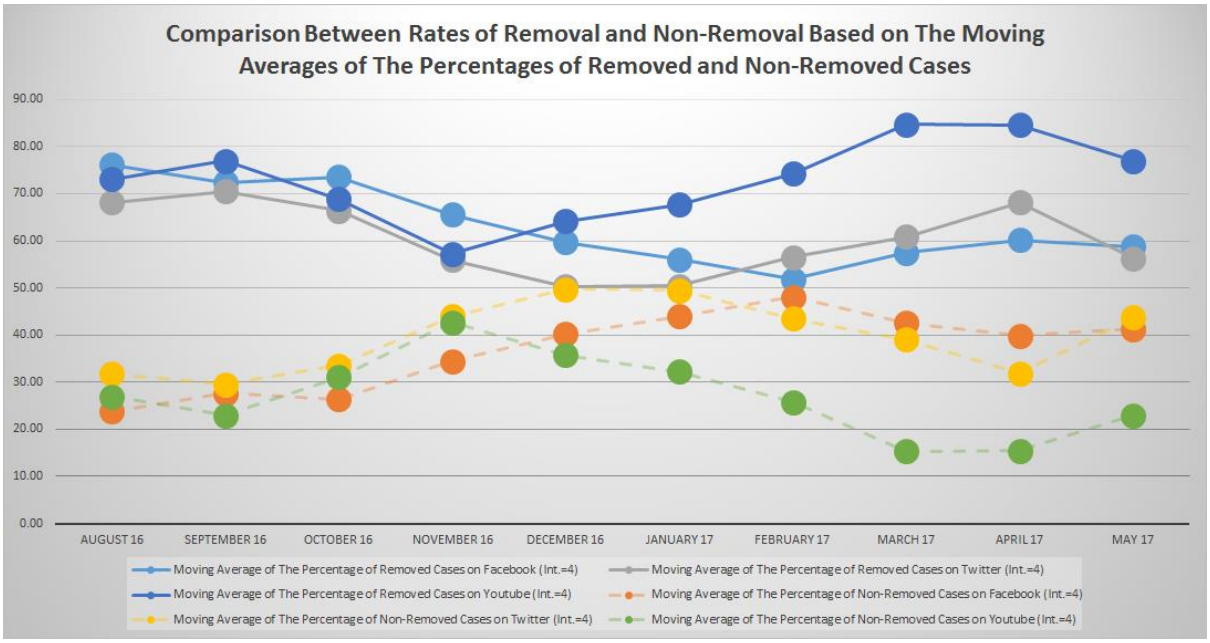


YouTube is far closer to Twitter than to Facebook when it comes to removal rates. Their numbers fluctuate in a very volatile manner, their lowest removal rates are between 40 and 50 per cent. However, they are the only major social media platform with an upward trend.



Although, this trend has to be taken with a pinch of salt. YouTube reached a very low number of removals by the end of last year and they came back from that during the first half of 2017, furthermore, their numbers - both on average and in an absolute sense - started dropping again, hence the moving average of their removed cases has been going down, just not enough yet to turn the trend around. It is yet to be seen whether this trend will continue or change.

On the final chart below, the moving averages of the removal rates of the three major social media companies are being compared, whilst their non-removal rates are also visualised. As one can see, with all the volatility in the monthly numbers and rates, and all the differences between the removal ratios of the different companies, their removal rates - on



average - do, somewhat, move together. They started out fairly high in the middle of 2016, dropped to very low numbers by the end of that year and from then on, they started going up again, at least until the past three months where they started going down or stagnating again. It is not clear to INACH and our partners why that is. The two monitoring exercises (discussed in detail later) carried out by the European Commission (EC) with the help of INACH members and other NGOs that were meant to monitor the adherence of social media companies to the code of conduct they had signed with the EC might have had an effect on removal rates. However, there is really no way to confirm a correlation or even connection between these numbers and the exercises. Yet, it is extremely interesting that the removal rates on these platforms move very similarly to one another. We hope to find out the reasons behind this fact in the future.

VII. A Debate Starter

The steady decline of anti-refugee hate, although this might change during the coming month, due to worrying news coming out of Italy and the EU's less than firm grip on the issue that signals a long and arduous process of negotiations that try to solve the issue and will ultimately probably fail, gives an opportunity to start a debate on effective approaches to achieving an open and inclusive society, and effectively countering cyber hate. As one can arguably extrapolate from INACH's data - just as in all other online phenomena -, "fads" and sudden trends can be observed in online hate speech. Offline societal, geopolitical, national and political issues are major and instant drivers behind cyber hate. Thus, hate types, such as anti-refugee hate, can appear suddenly and then slowly cool down and almost disappear as the drivers behind it slip into the background.

This might suggest that most people who create or share hate online are not radicalised dogmatic neo-Nazis or extremists. They most likely have not completely internalised the hatred that they spew against certain people or communities. If there is a new emerging

political issue that involves the proverbial “Other”, may that be refugees, Jews, Muslims or Romani people, they jump on the bandwagon and they project their insecurities and fears onto the “Other”. Hence, only a subset of the creators of cyber hate actually do what they do because of strong ideological convictions. Others, probably the majority, are “just” trapped in the eons-old “Us” versus “Them” mentality that - per definition - “Others” and securitises people and communities that dress differently, worship differently or differ culturally/physically from the members of the majority community. Thus, their “hatred” towards a certain group is most likely superficial, not internalised and abates rapidly if the online climate changes, at least towards that specific minority group.

This signals that to efficiently combat these issues, one must focus on the underlying causes, i.e. “Us” versus “Them” mentality, “othering”, securitisation and the fear of identity loss. It also means that first and foremost the notion of what it means to be the member of a certain European nation (e.g. being French, German, Spanish, etc.) has to change. Europeans cannot define themselves anymore through the colour of their skin and/or their religion. They should realise the excluding nature of these signifiers and come up with others that can really work as umbrellas that all can fit under. European values could be a great starting point. The respect for human rights, democracy, liberalism, secularism, etc. are all ideas and ideals that are inherently inclusive and independent of ethnic background, religious beliefs or skin colour. Obviously, immigrants and members of minority communities also have to subscribe to these ideas and - due to cultural differences - that will also be hard work on their part (and on the part of European societies). However, there are hardly any other options to create healthy and unfragmented European societies through integration that is not forced assimilation.

INACH does not argue that the above is an axiom within the field of cyber hate, but it is definitely something that is worth examining further, in order to be able to target and combat cyber hate in a more efficient way and develop counter techniques that incorporate these realisations. That is why we presented these ideas in a form of a debate starter chapter and that is why we welcome experts and other stakeholders in the field to think about these hypotheses. INACH will try to provide the needed space and time in the future to facilitate this debate.

VIII. The Fight Against Cyber Hate and Recommendations for the Future

The fight against cyber hate does not stop there. INACH and its partners did much more than collecting and analysing the data spread out above. Throughout the year, as well as organizing member and Steering Committee meetings among the project partners to ensure the constant information flow, INACH and its members organised multiple conferences to spread information and the project's findings to the world. On top of that, reports were published and workshops were organized. Moreover, within the framework of the Research – Report – Remove: Countering Cyber Hate Phenomena project, other activities such as INACH's social media presence and the establishment of a cyber hate database have also enhanced the work and reach of INACH, and helped the organization achieve its goals. To finish, regarding activities that were not part of the project, the monitoring exercise and its success will be discussed.

A) Our Activities and the Monitoring Exercise

One of the conferences that took place this year was the annual conference, titled "Taking Back The Digital Streets". NGOs, CSOs, representatives of the EC and social media companies all participated in the discussions on several aspects of cyber hate, as INACH, its members and other NGOs disseminated the data collected. Moreover, on April 19, 2016 INACH organized, together with its Israeli member ISCA, the first International Conference on Online Antisemitism, in Jerusalem. The conference, which was held on the premises of the Israeli Ministry of Foreign Affairs, had 130 participants from 21 countries, among them politicians, ambassadors, and representatives of NGOs. Sadly, the social media companies were unable to send their representatives. The conference generated a list of 17 excellent recommendations for all actors. The conference report can be found on the INACH website. INACH also participated in the OSCE Parallel Civil Society Conference 2016 on 6th and 7th of December in Hamburg and in (hate speech) trainings and seminars. INACH, and its partners also took part in an expert discussion titled "Combatting Hatred in the Social Media" at the civil society pre-conference of the OSCE conference in Berlin on the 19th and 20th October 2016, where recommendations for the OSCE participating states were developed.

Our partners also participated in their own events. For instance, in November, jugendschutz.net participated in the conference titled "Extremism on the internet" in Bratislava, organized by the Slovak Ministry of the Interior. Jugendschutz promoted INACH's activities and re-established contact with People Against Racism, a Slovakian NGO and a former INACH member that later joined INACH again. Furthermore, LICRA supervised and coordinated the launch of The No Hate Speech Movement in France on the 26th of January in Strasbourg. In July, LICRA also took part in the launching of private mediation proceedings for a legal summons to court against Facebook, Twitter and Google. Additionally, MCI participated in a conference organized by the National Police Corps on Hate Crimes in May 2017, titled "Hate Crimes and Discrimination" that took place in the Law School of Málaga in March. ZARA Training, ZARA's affiliate, that is tasked with prevention work, held a training consisting of two workshops on cyber hate in May 2017, one

at a vocational school as part of its peer education project and one open workshop for adults at an education centre for adults. Furthermore, in March 2017, ZARA participated in their annual press conference for the presentation of its Racism Report of 2016, as well as introducing its counteract platform launched in January 2017 that can be accessed at: http://www.counteract.or.at/. This platform provides information, tools, and instructions to help tackle hate on the Internet effectively. The web page also informs about initiatives, campaigns, education and research on the topic of cyber hate.

INACH and its partners also produced and published reports such as the one from August 2016, titled “Kick them back into the Sea”. This report gives an overview of the rise of cyber hate related to the so-called "refugee crisis" in six European countries. The report can be found on INACH’s website. Regarding our partners’ accomplishments, jugendschutz.net published an article about INACH in the magazine called "Gegen Vergessen – Für Demokratie", the newspaper of one of our German INACH members, and MCI published a report on hate speech titled “Apuntes Cívicos” in December and issued a new RAXEN Report on Hate Speech in May 2017, titled “Contra el Discurso de Odio y la Intolerancia”.

On other matters, INACH launched its new and completely restyled website, paired with its Facebook page and Twitter account. We have been posting on these platforms ever since then on a regular basis, providing snippets of data, raising awareness and publishing all our reports. This activity helped to boost our public reach and ability to disseminate our findings.

Another way INACH is working on solutions to try and solve issues related to cyber hate, is by developing an international online cyber hate database and complaints system. We explored existing content guidelines, researched legal provisions and managed to define standards for content guidelines related to online hate speech. We also developed definitions and standards for recognising and documenting cyber hate. Furthermore, these standards and definitions were agreed upon by all project partners, which is a major achievement in a field that is as contested as hate speech. After finishing the theoretical background, INACH started working on the technical side of the complaints system and database. This consists of three separate but interconnected segments. The first one is the complaints form that is used by the public to report instances of cyber hate to the member organisations of INACH. This form forwards these complaints directly to the international online complaints system, specifically to the respective organisation that is responsible for complaints for the specific country that the complaint came from. The second one is the complaints system. In the complaints system, the complaints officers can record and administer everything that is needed to notify social media platforms, register complaints, categorise them, get these instances of cyber hate removed and then close these cases. The third segment is the database where all results of the previously described procedures end up. This vast amount of data, supplemented with studies, event timelines, news articles, legal documents, reports, studies, journal articles and other useful information pertaining to online hate speech provide one of the most extensive data mining opportunities within the field.

In addition, the monitoring exercise was another stepping stone for INACH and its fight against cyber hate. On the 31 May 2016, the European Commission with Facebook, Microsoft, Twitter and YouTube ("the IT Companies ") signed a "Code of conduct on countering illegal hate speech online". The main commitments were: a) The IT Companies to have in place clear and effective processes to review notifications regarding illegal hate speech on their services so they can remove or disable access to such content. The IT Companies to have in place Rules or Community Guidelines clarifying that the prohibit the promotion of incitement to violence and hateful conduct. b) Upon receipt of a valid removal notification, the IT Companies to review such requests against their rules and community guidelines and, where necessary, national laws transposing the Framework Decision 2008/913/JHA, with dedicated teams reviewing requests. c) The IT Companies to review most valid notifications for removal of illegal hate speech in less than 24 hours and remove or disable access to such content, if necessary. The IT Companies and the European Commission agreed to assess the public commitments in the code of conduct on a regular basis, including their impact.

To ensure an effective measuring of progress, the Commission's sub-group on countering hate speech online agreed, on 5 October 2016, on a common methodology to assess the reactions of IT Companies upon notification of illegal hate speech. It was also agreed that the preliminary results of this monitoring exercise would be reported to the member states, the IT Companies and to civil society organizations within the framework of the High-Level Group on combating Racism, Xenophobia and other forms of intolerance. INACH was instrumental in aiding this process and in helping to create the methodology. Furthermore, most of INACH's EU-based members provided the necessary monitoring data. For 6 weeks, 12 organisations based in 9 different Member States applied the common methodology. The organisations notified alleged illegal hate speech online (as defined in national criminal codes transposing the Framework Decision) to the IT Companies and used a commonly agreed template to record, when possible, the rates and timings of takedowns in response to the notifications. The monitoring exercise was a continuous process. The collected data constituted a baseline and a first valuable indication of the situation. The fairly abysmal results of the first monitoring exercise can be found here.

A second monitoring cycle was carried out during 2017 to observe trends. INACH had a pivotal role again in the tweaking of the methodology and the development of the data collection tool. Regarding the results, they were published on June 1st, 2017 by the European Commission. In this exercise, 2575 notifications were submitted to the IT companies taking part in the Code of Conduct, 31 civil society organisations and 3 national authorities, located in 24 EU countries took part and Facebook received the largest amount of notifications (1273 cases), followed by YouTube (658 cases) and Twitter (644 cases). Confirming the trends that we discussed in this report, those results confirmed a predominance of hatred against migrants and refugees (with 17.8 per cent for xenophobia including anti-migrant hatred), as well as anti-Muslim hate (17.7%) which were also among INACH's top hate types. Regarding removal rates, it was observed that 59.1 per cent of the illegal hate speech cases brought to the social media were removed by them, which represents an increase compared to

last year's monitoring exercise where the rate was 28.2 per cent. Concerning non-removal rates, Twitter did the worst by only removing 37.5 per cent of the content, followed by YouTube with 66 per cent removal rate and finally Facebook with 66,5 per cent removal rate. Regarding the 24h response rate, it went up from 40 per cent to 51.4 per cent in a year. However, it was observed that cases brought by people other than trusted flaggers were not as well received and successful in their removal rate as those flaggers, and this is one of the main issues to be taken into account for further improvements. Indeed, 56.5 per cent of the notifications made using channels available to general users led to the removal of the notified content, while a higher removal rate of 65.6 per cent was recorded for notifications made using the trusted flaggers/reporters channel. Moreover, Twitter provided feedback to 68.9 per cent of notifications made using the trusted flaggers' channel, but only gave feedback to 13.4 per cent of those notifications made by general users. For YouTube, the corresponding figures were 35.5 per cent and 15.6 per cent respectively. Moreover, even though the ratio of response within the first 24 hours went up to 51.4 per cent, a more than 10 per cent increase, responses within 48 hours only reached 20.7 per cent. This means that companies responded within 48 hours in around 72 per cent of the cases, a more than 10 per cent drop from the previous monitoring exercise where this number was 83 per cent. The official outcome of the second monitoring exercise can be found here.

B) Recommendations

Even though INACH and its partners have made essential steps forward in the fight against cyber hate over the year, much remains to be done. Three main issues that need remedy stand out. Firstly, the discrepancy between what material is legal and illegal is a problem. Indeed, too much cyber hate material, considered hate speech by the public or by INACH members, remain in the legal realm, which hinders the efficacy of the work of NGOs like INACH. There is not one clear universal definition of hate speech, hence there is no EU wide consensus. Furthermore, the field's clash with free speech makes matters increasingly difficult. Nonetheless, transnational institutions and governments have produced a number of treaties and legislation to remedy the issue, and those can be found in the report titled "Legislation related to cyber hate" on our website. However, efforts still have to be made here and laws should be tightened regarding cyber hate, and consensus should be found.

The same goes for the content guidelines of the social media platforms we monitored. Here, the inconsistency between what is being removed and what is not shows that we are in dire need of a long-lasting solution. Those major differences in removal rates imply that social media companies interpret their own rules and guidelines subjectively and arbitrarily, making the work of NGOs and other organizations much harder. Removal rates are highly influenced by the amount of complaints received, and by who the complainer is. If it is an authority or a very well established local NGO, or other civil society organization that is a trusted reporter or flagger, it is much more likely that the hateful content will be removed, as it was the case for the monitoring exercise discussed above; just like when a lot of people complain about a certain content. This should not be the case. Illegal content and content that violates the guidelines should be removed globally and universally, irrespectively of the number of

complainers or who the flagger is. Otherwise, extremist groups and people who hold extreme ideas and ideologies are enabled and highly illegal, violent, hateful and vile content is left online for months without any real explanation from social media giants, whilst minor and even benign infractions are removed within hours. This attitude and the companies' modus operandi must change, if we are ever to have an online community that respects the human rights of all its members. This especially goes for the three major social media companies, namely Facebook, YouTube and Twitter, that must recognise that their unmatched reach that affects enormous amount of lives bestows upon them a social responsibility that they need to live up to. Our study on these issues can be found here. Additionally, another issue that illustrates the inconsistencies of removal rates, is the issue of the removal of hate speech content such as Holocaust denial. In May, *The Guardian* leaked information about Facebook's internal instructions on hate speech and other issues. In countries where there is no legislation, or unenforced legislation, Facebook takes no action against Holocaust denial. This was the same for migrants, refugees and asylum seekers who were deemed as "quasi-protected category", hence not receiving the protections given to other vulnerable groups. This simply confirms that there is a real issue there, which needs to be remedied, namely in the form of consistency and common sense, in order to have a cleaner internet.

Thirdly, a major issue arose concerning the monitoring exercise. There was an undeniable bias that took place due to the fact that social media companies were informed about many details of the exercise, such as when it was going to take place and who was involved. For the next monitoring exercise, and to avoid this bias, it would then be advised not to inform the social media beforehand about the details concerning when the exercise will be taking place. Moreover, the ongoing exercise should be masked somehow, so the companies will not be able to recognise that it is going on. This could be done by using low level monitoring for a longer period of time or by carrying out the exercise in a rolling manner, where NGOs do not start it at the same time and end it at the same time, but spread it out for a longer period and start one after another. This way the elevated activity could be a bit more hidden.

To conclude, cyber hate is a reality that grows more destructive each day. Due to cyber hate's omnipresence on social media, making it cheap and easy for extremist to spread their message, the general public sees and consumes cyber hate on a regular basis, leading racism, antisemitism and other types of hate to become normalised. People simply become desensitized to these issues and accept cyber hate as truths. This was seen in our trends, where for instance, Muslims and/or refugees become linked to criminality. Although cyber hate is not the only driver behind this rise such discriminatory sentiment, it enhances the normalisation and acceptance of such views. That is why it is pivotal for stakeholders such as authorities, EU institutions and European governments to recognize the dangers presented by cyber hate and develop policies such as the ones we mention above to counter it.

IX. References

The data (both qualitative and quantitative) in this report was collected from INACH members and project partners, specifically for the purposes of Project Research - Report - Remove: Countering Cyber Hate Phenomena. These partners are: **jugendschutz.net (Germany), the Inter-Federal Centre For Equal Opportunities and Opposition to Racism (now Unia) (Belgium), the Zivilcourage und Anti-Rassismus-Arbeit (Austria), the Ligue Internationale Contre le Racisme et l'Antisémitisme (France), the Movimiento contra la Intolerancia (Spain), the Meldpunt Discriminatie Internet (Netherlands) and the Meldpunt Internet Discriminatie (Netherlands)**.