BEYOND SOCIAL

a network and sentiment analysis of Europeans' conversations about the Future of Internet

HER

PLUSVALUE

Beyond Social

a network and sentiment analysis of Europeans'

conversations about the Future of Internet

This work was done by Fiorenza Lipparini, Oriana Persico, Salvatore Iaconesi, Guido Romeo, Emanuele Rizzardi, Eugenio De Matteis. We had great fun designing and discussing this piece of research, as well as lots of sleepless nights validating tons of data and browsing through network graphs to understand their meaning. We would be very happy to hear from everybody interested in these topics!

Index

Foreword

This report offers a first reading of the data gathered between **November 2016 and April 2017** by listening to public conversations on the **Future of the Internet** (FoI) on major social networks (*Facebook*, *Twitter*, *Instagram*).

The aim of this research was to better understand stakeholders' thoughts, expectations, desires, concerns, visions and imaginations based on their expressions when talking about internet technologies and their impact on our lives, societies and the economy.

The research brings together distinct types of stakeholders, including citizens, experts, activists, businesses, civil society organizations, public institutions, based all across Europe and beyond, and speaking 54 different languages.

This is an incredible richness and variety.

This same richness, however, is what makes nearly impossible to extract a single meaning from the data. Indeed, trying to infer a single interpretation, or meaning, from all of this richness and variety would be a great loss, as the value of a study such as this one is to reflect complexity, and to be able to welcome, understand, value and, to some extent, manage it.

For this reason, in the following pages we will try to convey a broad spectrum of interpretations, based on what we detected across different communities, territories, cultures, contexts and types of subjects.

All the data we collected and processed are accessible in form of usable, accessible, Open (Big) Data sets, so that everyone can use it for research, policy, design, art and to understand where we are going, as a deeply connected society.

In the following sections we will:

- **outline** synthetic results, to provide accessible and concise indications to be shared, used, disseminated and, hopefully, further tested and elaborated.
- **detail** divergences and variations, to highlight how themes and issues are confronted across diverse cultures, communities and according to different approaches;
- **highlight** issues, to pinpoint the most controversial and urgent topics to be addressed by policy-makers and stakeholders;
- **provide** meaningful data sources to boost further research and sharing.

We invite everyone who will use the data to share it again using similar, permissive, open licenses, and to get in touch with us to compare results and to collect them in a single place, for the benefit of all stakeholders.

1. Introduction

Who is talking about the Future of Internet (FoI) on social networks? In which terms? And why?

These are very challenging questions.

Indeed, some people talk about the FoI because of their profession, since they believe that a better, different, more advanced Internet could provide opportunities for their businesses and organizations. Some – not necessarily different people – talk about the FoI because they are concerned about it, they feel there is something fundamentally wrong with the way things are evolving, for instance in terms of privacy, security, democratic processes which might be undermined by opaque algorithms and big data based profiling, and of course in terms of access and human rights. Interestingly, the adjective most often associated to these concerns on social media is “not smart”, which might suggest that in order to be smart, next generation internet technologies also need to be secure, transparent, respectful of privacy and inclusive. And indeed, many conversations are about the internet intended as the enabling space where not only innovation, but social transformation will take place in the near future, if we are able to build and foster certain features, strategies, approaches, and collaborative / participative practices. Such a space would be the ideal playfield where like-minded people and organisations could successfully address global issues such as war, poverty, access to healthcare, clean energy and so on.

For many users, and particularly younger audiences, conversations about the FoI are about having fun, while for others we cannot really speak about conversations, but more correctly about the constant sharing of news on this topic, which sometimes generates more engaged debate.

Of course, it is not always easy to tell when people are talking about the FoI and its socio-economic consequences. For instance, are we talking about the FoI when complaining about broadband speed? Or when wondering how exactly our social network provider was able to show us that advertising banner which was in particular thematic sync with that private message we just sent? Or when chatting with friends worrying about what is going to happen in the next global cyberwar? What about sharing the weird news about that Smart Refrigerator that has been hacked (and now pirates from some weird country know all about your eating habits?).

Are we talking about the FoI when talking about anything that, at every moment, generates data or requires network connectivity of some sort? Or, going in the opposite direction, what exactly, today, does not involve data and connectivity? Is there such a thing anymore? And are we still talking about the FoI when commenting precisely about such a thing, if it exists?

When we started planning and designing our social media research about the FoI, we were asking ourselves all these questions, and many more. Our aim was to discover what European citizens were thinking and discussing spontaneously in relation to this topic, without the bias of

questionnaires, surveys, communication campaigns, in their daily lives, without us asking.

We wanted to know if there was a way to find out whether the "Future of the Internet" was something on people's minds, to what extent, and why. For this reason, we split the research in two phases.

In the first phase, we searched for content on major social networks (Facebook, Twitter, Instagram), compiling a list of terms which were relevant from a FoI perspective, in 54 languages. We used combinations of words and terms to identify those expressions (posts and tweets, for example) which referred to internet technologies in terms of future expectations, desires, needs, fears, visions. For example, posts like: "my online banking account needs to be more secure than it is now", or "I would expect my wearable device to protect my data".

As we collected data according to this first strategy, we prepared the second phase: we used several techniques (mostly Natural Language Processing and Machine Learning) to gather advanced information about the ways in which subjects expressed about the FoI, for instance in terms of languages and forms, channels, groups, words used by different types of subjects in different situations and contexts.

In fact, the first phase allowed us, starting from our assumptions on key words and combinations of words likely to be used when talking about the FoI, to better understand how people where really expressing about this topic, using what words, languages, forms and patterns. Based on this knowledge, we recalibrated our collection deck, and started the second phase, monitoring conversations in light of what we knew where the most recurrent relevant words, languages, forms and patterns. This of course also allowed us to produce networks.

These networks of topics and scenarios are, in fact, the most outstanding result of this research. We have collected and processed relational networks which indicate which topics and sub-topics go together, in which situations and for what types of people.

We can now better understand to what extent - in the social network world - European citizens talk, for instance, about robots, and what do they associate this topic with; for instance, jobs, and which emotions they associate to this association (maybe they're afraid that robots will steal their jobs?).

We can say when, and how much, people talk about the Internet of Things (IoT) and Cyber Security. Or, in other cases, about business opportunities. And if they're scared, nervous, excited or curious about all these topics and issues.

In the following pages we will, firstly, provide some context about the information collection process. Secondly, we will dive into these networks, with a view of progressively answer our questions about how people see the future of the internet in complex and - hopefully - useful ways. In the annexes, a full methodological section is available, including references to where to access the data sets, how to play with it, and how to share the results.

Let's begin.

2. Overview

2.1. Who's talking about the FoI?

We have collected **669.734** messages generated from **355.451** users, between **November 10, 2016** and **April 30, 2017**.

Source	Number of content
FB Groups	14.129
Twitter	655.605

Messages were written in **54 languages**.

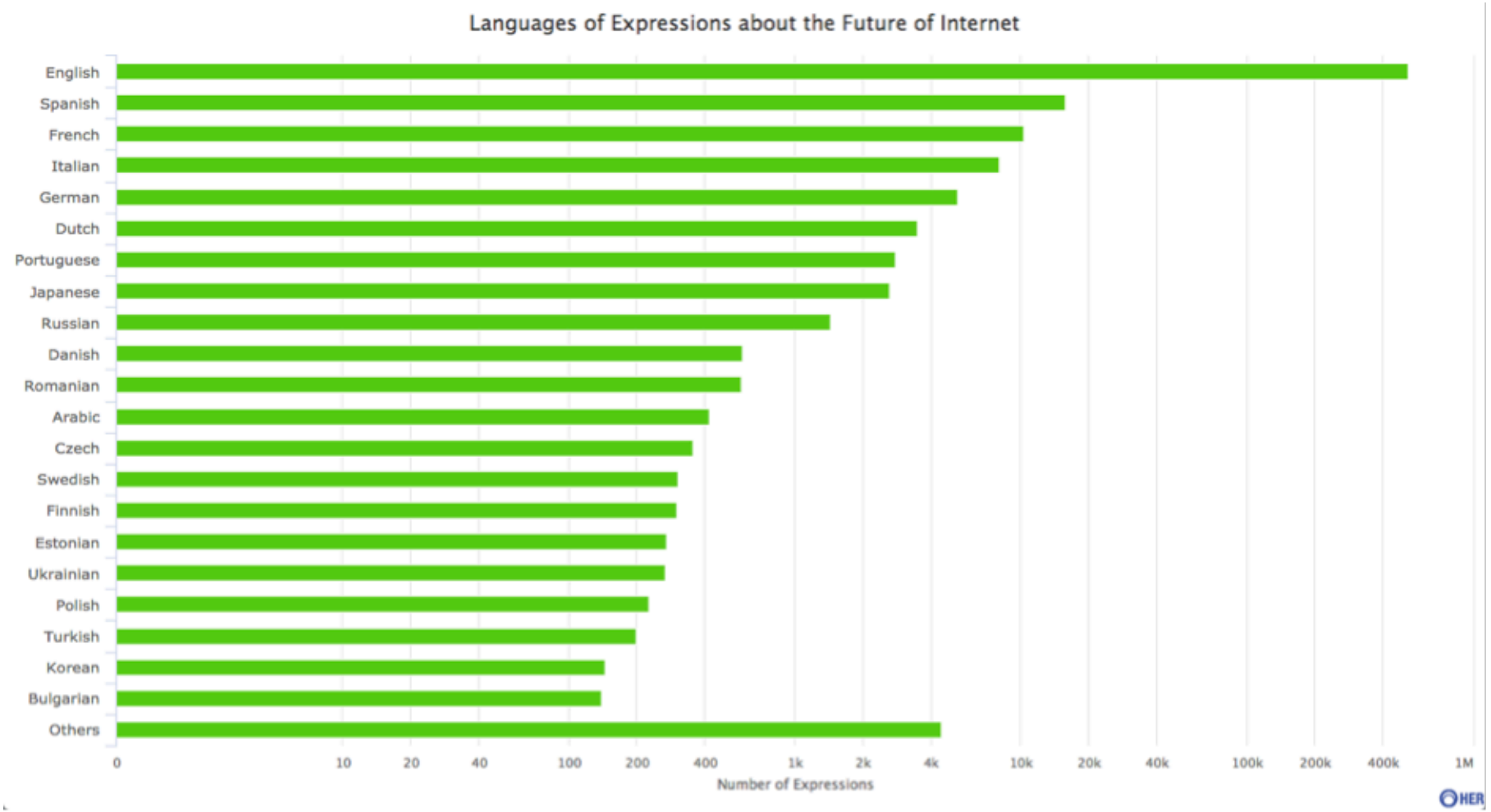
The distribution of languages (shown in the graph using a logarithmic scale) clearly shows the predominance of the English language: more than 400 thousand messages, compared to the around 10 thousand in Spanish, French and Italian and the around 5,000 in German and Dutch.

To better understand to what extent people using social media were interested into the future of the internet, we monitored over the same period of time, using the same techniques ad on the same media, how many people were speaking about what we perceived as trending topics, such as Brexit and Soccer. Not surprisingly, conversations about soccer were incomparably more frequent than conversations about the future of the internet, however, the situation changes if we use Brexit as a comparison, denoting a broad interest in the theme.

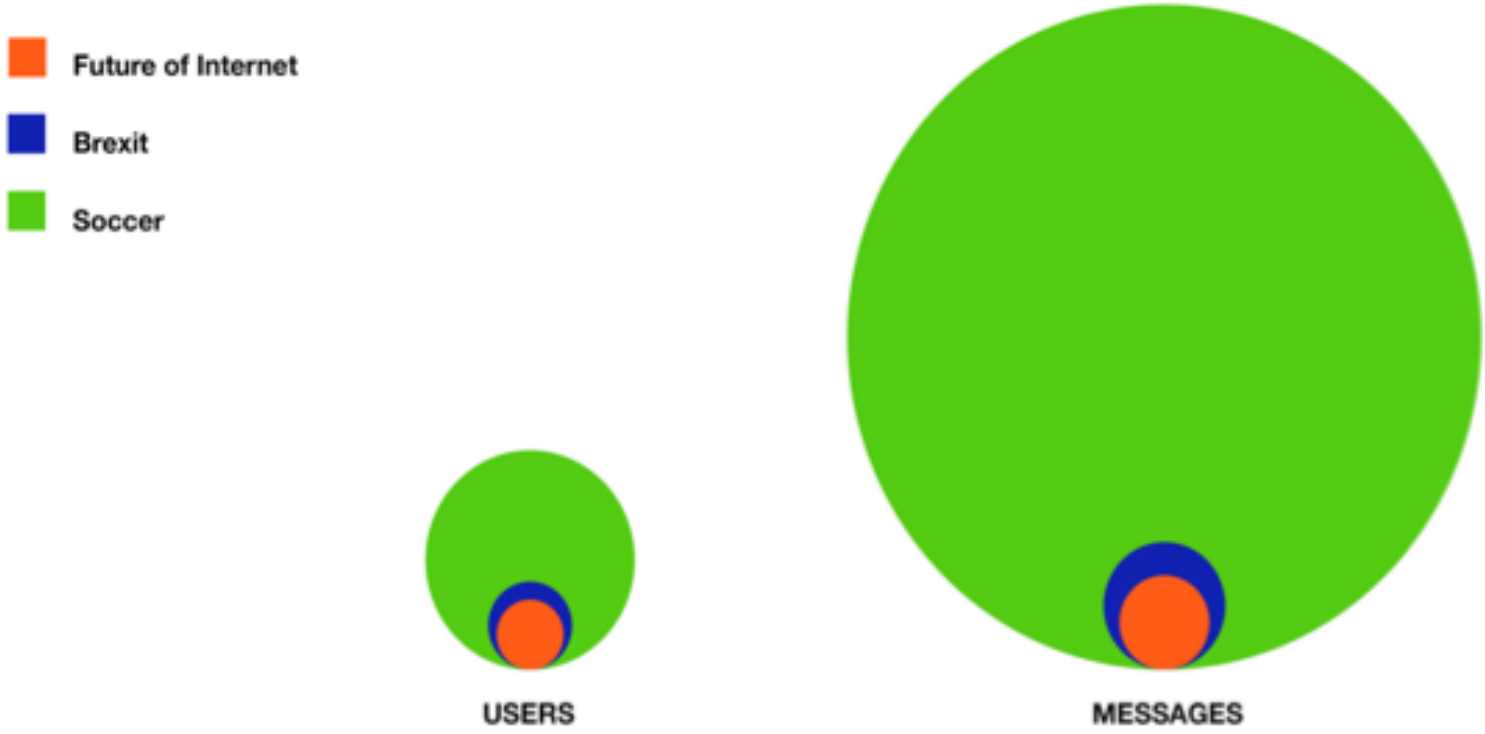
But who was interested in the FoI? As a first thing, we tried to understand if users were really humans, and if they were individuals or organisations. For this we used available techniques (see methodology) to analyse statistical parameters such as frequency/patterns of posts, regularities in mentioning tags and other elements, so as to infer with a statistical probability the types of users which did or did not post FoI expressions. For each of these parameters we set specific ranges which we used to classify the different types of users: *for example, organizations, in this sense, are similar to bots, but have lower volumes of posts, are less strict in the forms of content they emit, and have a more varied type of response pattern.*

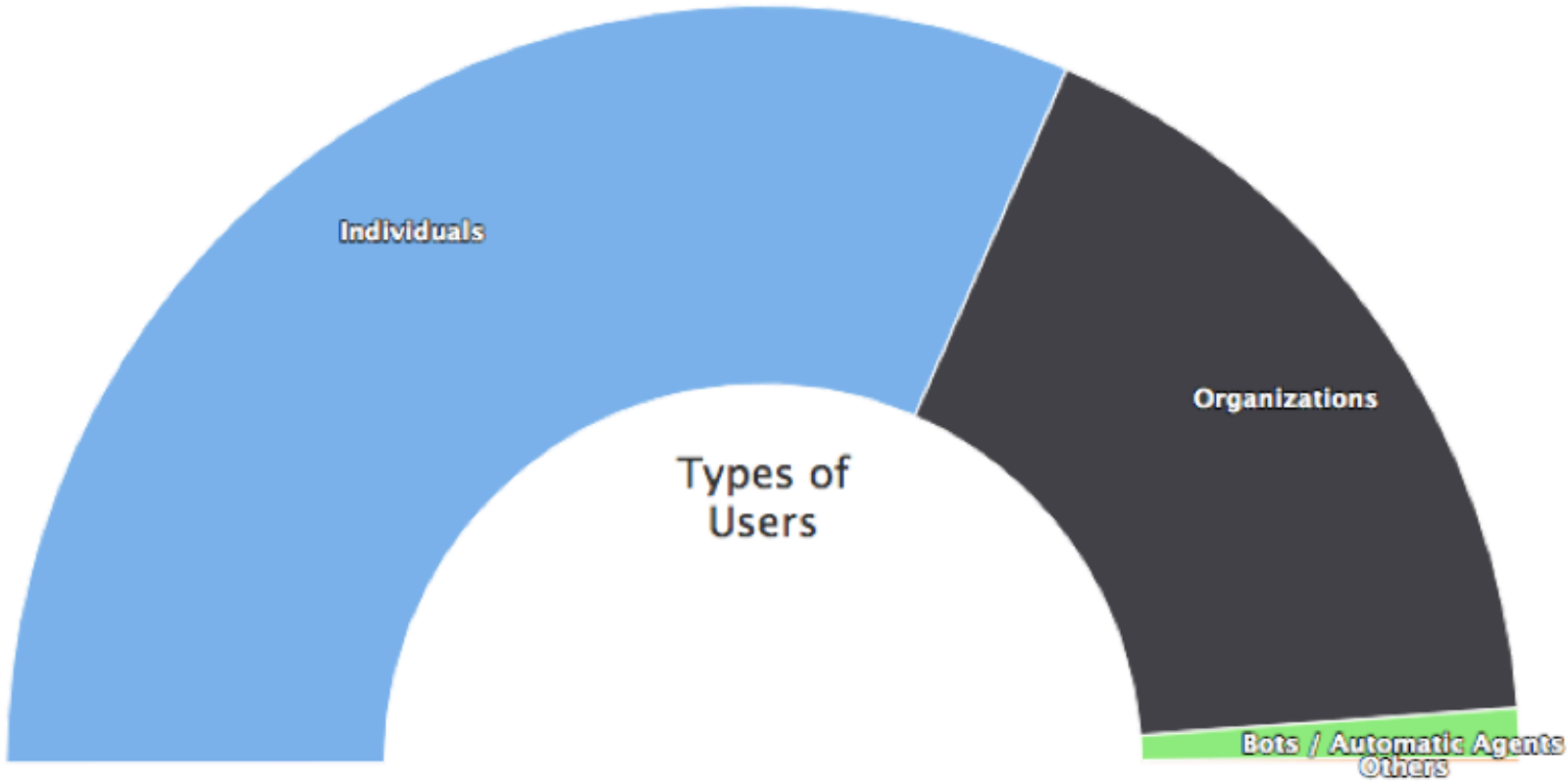
What we found (Graph 1-3) is that while bots and automatic agents represented only about 2% of users, they produced nearly 40% of the messages. Then we tried to distinguish between individual and organisational accounts, and found that organisations accounted for about the 34% of total users and 33% of messages produced, leaving a 63% of individual users responsible for the 27,5% of the messages retrieved

Graph 2-1 The languages in which expressions have been collected.



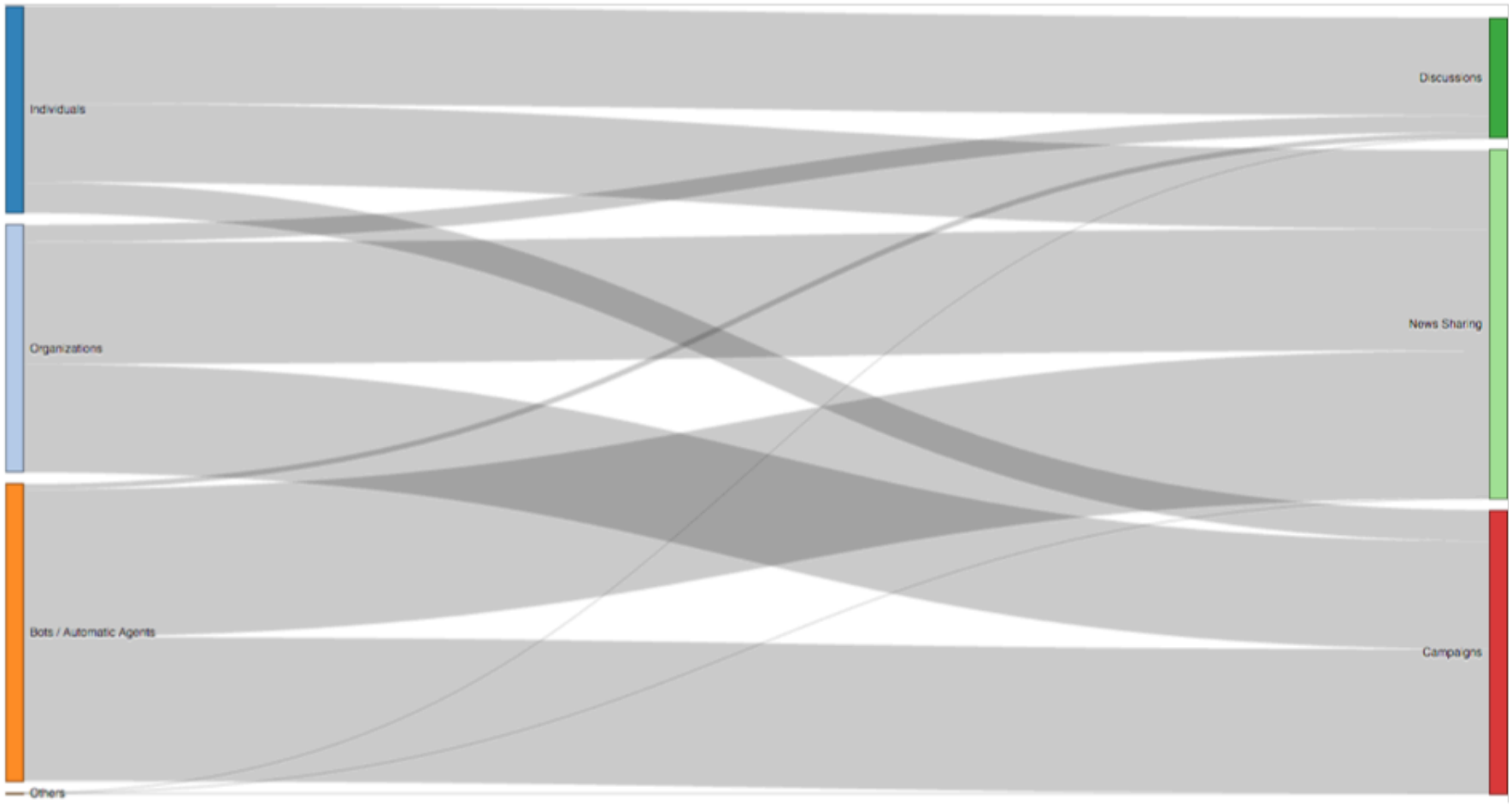
Graph 2-2 Comparison between the numerosity of expression across themes.



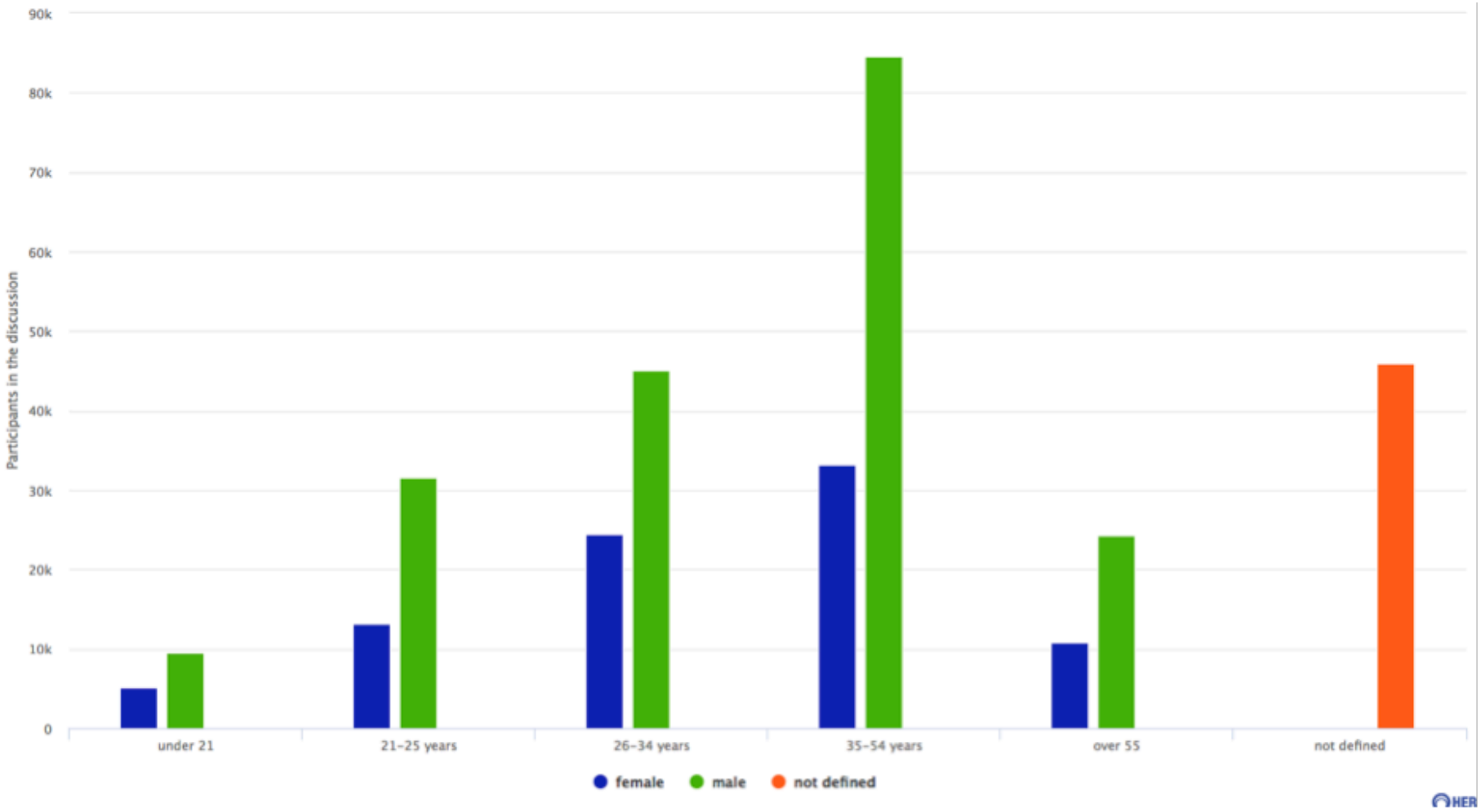


Graph 2-3 Humans and non-humans talking about FoI.

Graph 2-4 The flow of conversations on FoI.



Graph 2-5 The ages and genders detected in the subjects expressing about the Future of Internet.



What is more interesting though, is to see how different the messages are based on the nature of their originators.

As we can see in Graph 1-4, the quantities of messages generated by each type of user are very different, while they appear in very different and specific types of conversations[1]: in fact, while **individuals focus on discussions, news sharing and discussing practices**, **organizations and bots practically engage only in systematic and voluminous news sharing patterns, and in their own campaigns**. This also gives us a measure about how many of the conversations we collected are pertinent to actual discussions and how many are instead campaigns or other similar activities[2].

We also analysed the gender and ages of the subjects who expressed about the FoI (see the methodological section to understand how, and with which precision).

Interestingly, as seen in Graph 1-5, far more men than women (70% vs 30%) are discussing about the FoI online. This of course is due to a variety of factors, has showed by an increasingly broad body of literature. Unfortunately, despite increased attention by European and national policy makers, the gender gap in terms of ICT access, skills as well as in terms of (self-)confidence and interest into the subject is not narrowing. This is confirmed by the EU Digital Scoreboard 2017[3], according to which, in spite of the fact that more women than men acquired ICT skills in formal education, and that the difference in terms of access is minimal (87,4% of EU men vs 85,6% of EU use the internet every day), only the 3,5% of EU women have written a computer programme (vs 9,6 of men), while men graduated in science and technology are twice as many as women. If we look at the Special Eurobarometer 460, men are more likely to agree for all aspects asked about than women (they are more positive about ICT impacts on their lives, society and the economy, they are more informed and more optimistic about the current and future role of social media, AI and robots, they feel more skilled in using ICT in current and future working and social contexts, etc.) with one single exception: indeed, “there is no difference between men and women when it comes to feeling sufficiently skilled in the use of digital technologies to do their job”. This suggests that the workplace could be a good starting place to engage women in a debate around the FoI, however, further research to understand gender differences in terms of language, topics and patterns adopted when expressing

[1] Here we have described as “News Sharing” those conversations which substantially focus on someone sharing a link to a news item, “Campaigns” those conversations which revolve around the systematic and prevalent use of a single recognizable element (for example an hashtag) and “Discussions” all the rest.

[2] This does not mean that it is not important or interesting that a company, for example, creates a certain campaign on a certain FoI topic, using a certain tone of voice, emotional approach etc. On the contrary: it is of vital importance to understand these types of actions, as well as who is investing in automatic (ie: bots) social network activity on certain issues and all related matters. It gives a better understanding of the interest involved, of communication strategies, of communities involved and other interesting elements.

[3] http://digital-agenda-data.eu/datasets/digital_agenda_scoreboard_key_indicators/visualizations

about internet technologies would be needed. Concerning age groups, the **35-54 age range** is the one that expressed the most about FoI (35-54 years, male: 84606, female: 33218). It is remarkable and worth noticing how younger generations (*under 21*) barely express at all, in comparison. Even in this case, there might be several explanations. Firstly, we are witnessing a progressive shift of younger generations towards different platforms for their public and private expressions and interactions. Among these are other social interaction and media platforms such as WhatsApp, Snapchat, Telegram, and other minor ones, which have hybrid characteristics

falling between instant messaging platforms and social media. As new platforms appear on the market, younger generations move fluidly among them according to a nomadic, unstable, profile. Of course, platforms such as WhatsApp, for example, have established a strong presence and user base, but the overall scenario is far from static and very fragmented and dynamic. One characteristic of these platforms is their “closureness” and strong particularization. Services like WhatsApp and Snapchat, for example, are constituted through the aggregation of myriads of social microcosms. This raises several worrying issues, for instance the likely emergence of social, knowledge and information bubbles, and the quasi impossibility for large, shared, public discussions to emerge.

This fact alone could account for the limited contribution of younger generations to the conversations analysed by the network and sentiment analysis. But there’s more. Indeed, large numbers of young people still actively use the social media services which have been screened in this analysis: according to the Standard Eurobarometer 84 (Autumn 2015), 77% of Europeans aged 15-24 use social media every day, and another 13% use it at least once a week. We therefore need further hypothesises to account for their absence from our sample. Looking at their online behaviours the following theses emerge, including:

- different linguistic styles used for expressions;
- different topics, scopes and objectives of social media communication; and
- different approaches to usage of technological systems.

In summary, it seems that younger generations use different vocabularies, and have different linguistic styles and behavioural patterns to express about the FoI compared to our adult sample. Which ones? The answers are found by analysing their scopes and objectives for social media communication, which are more playful and fluid, if compared to the ones of older generations. On top of that, younger generations tend to have a more utilitarian vision of the platforms they use: they are more vocal about the malfunction of certain functionalities, than about the technological, social, economic, or political implications of the technologies they use. If not stimulated, they appear to be more “users” than critical discussants. This, if united with the changing style and behaviour of social media expression, which is becoming generally shorter in form and simpler in content architecture, provides useful insights about the need to elaborate specific communication

strategies – for instance by using design, arts, or gamification processes and, in any case, giving serious considerations to the visual, linguistic and iconic styles to be used - if a wider participation of younger generations is desired.

2.2. Topics and Emotions. What are people talking about?

It is very interesting to see what are the most discussed topics and the emotional expressions related to each of them.

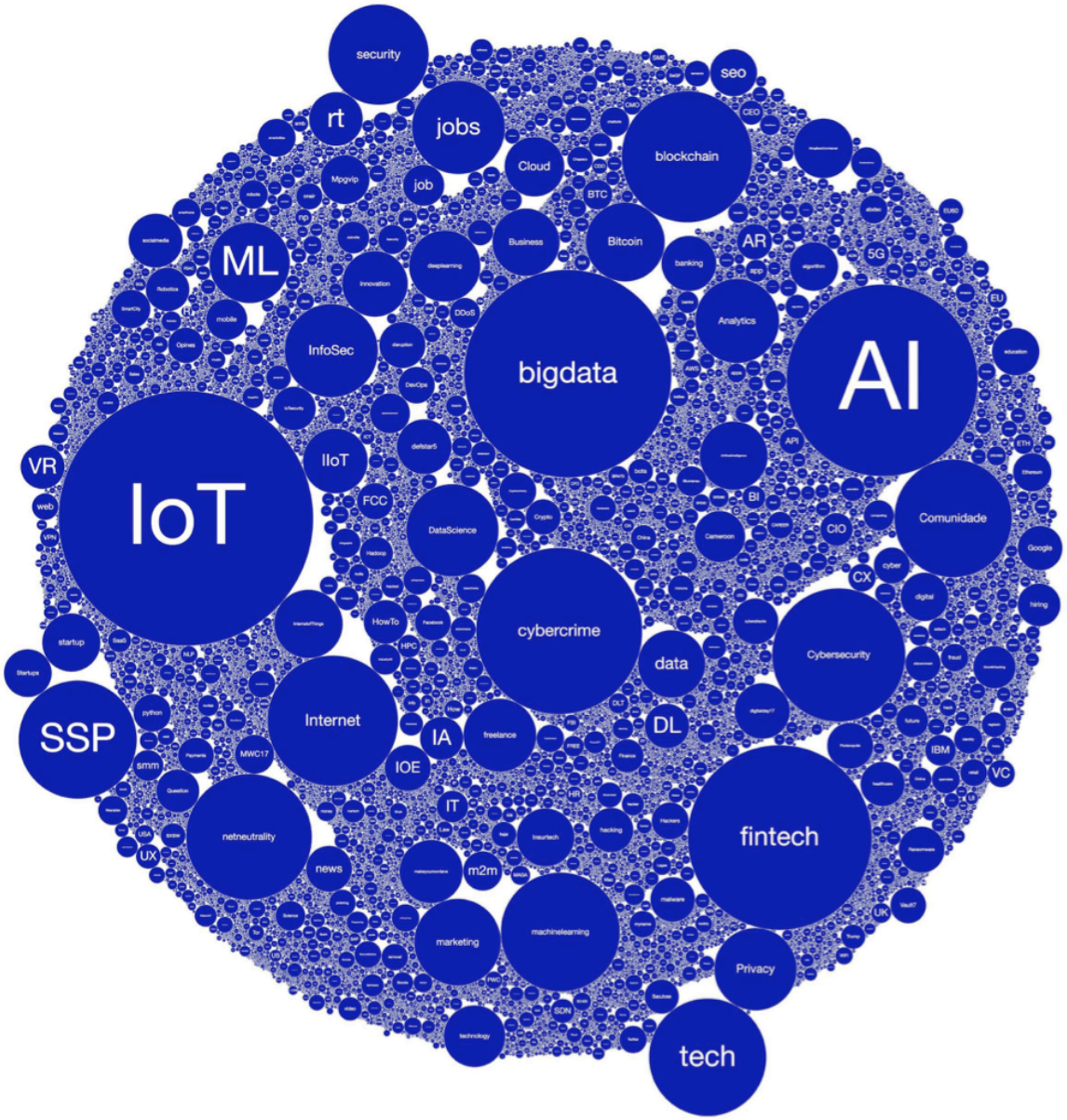
If we look at the topic cloud (Graph 2-6) it is immediately apparent how **a limited set of topics is the principal focus of discussions on the FoI**: IoT, AI, BigData, Fintech, Cybercrime, Cybersecurity, Machine Learning, Blockchain, Net Neutrality, Privacy and a few more.

But how do people feel about these topics? In our model, emotions are classified according to their value of Comfort/Discomfort and Arousal, according to the Circumplex Model of Affect (see methodology[4]). This means that each emotional state corresponds to a level of Comfort and Arousal/Energy, as shown in the image below.

Graph 2–7 and 2–8, representing the emotional distribution for single posts on social media, show how the largest clusters are to be found in the positive-comfort/positive-energy, positive-comfort/negative-energy, and negative-comfort/positive-energy quadrants.

The positive-comfort/negative-energy quadrant, corresponds to emotional expressions of confidence, seriousness, contemplation, or of being impressed. Discussions in this cluster tend to be “normal”, discreet discussions, in which fights and arguments are rare and communication is mostly "information", with subjects calmly presenting their points of views.

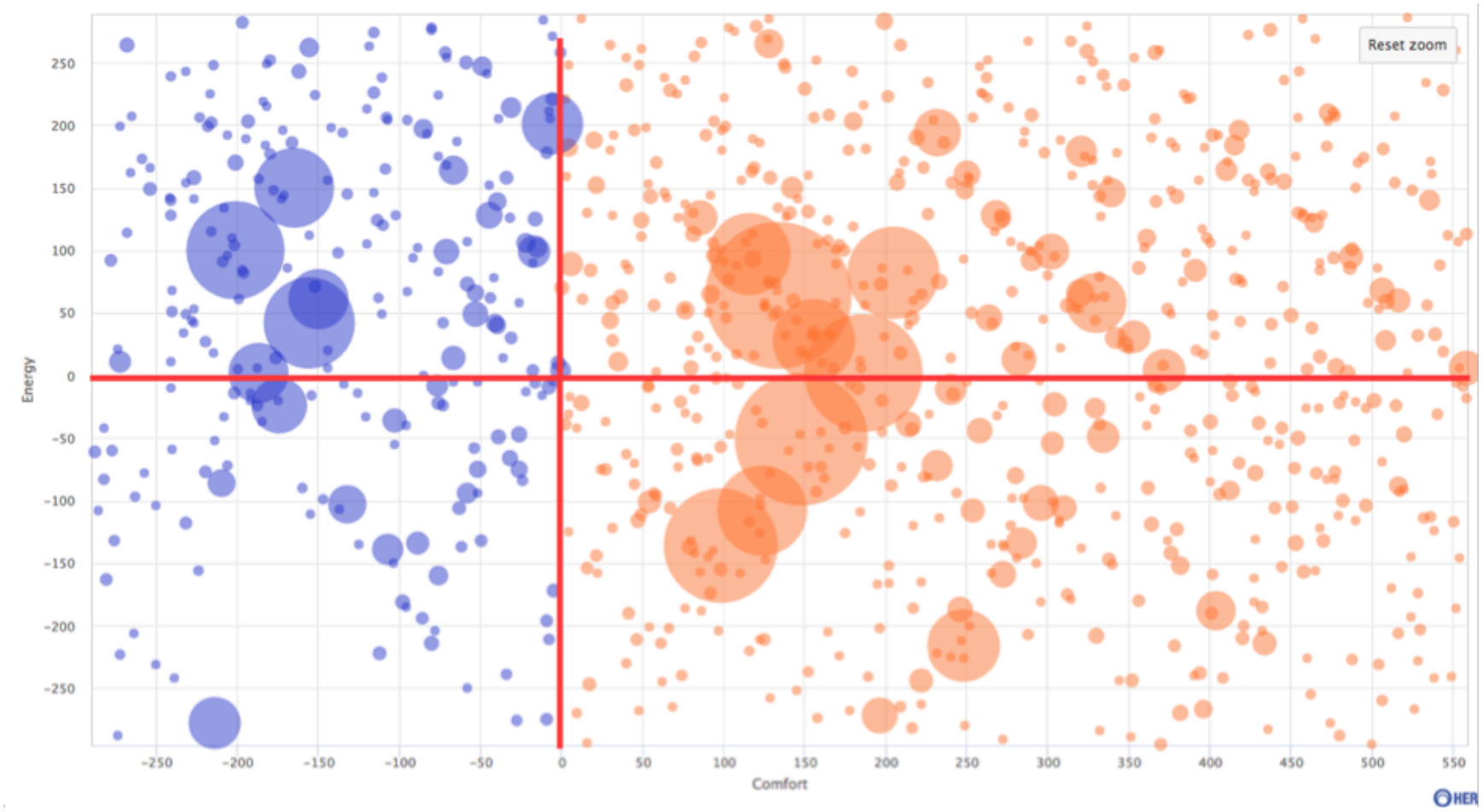
[4] The model uses dimensions of High/Low Activation and Pleasant/Unpleasant, which are maybe the most direct to map in semantics (eg: in words and phrases). We started with emotional derivations of Wordnet (such as WordnetAffect and others), validated for integration and harmonization through thousands of tests on Mechanical Turk, and then, with the department of linguistics at Yale with which we build the machine learning that we still use to augment the vocabulary. Basically: we use machine learning to discover patterns in language in which words are systematically used in contexts defined by other words for which we already know their characterization in terms of affect. For example, if a certain word or phrase systematically appears (we set thresholds, for example of 100 times) in a context which is strongly typed as X-activation and Y-pleasure, we flag it because it may represent a good candidate for this kind of emotional classification.

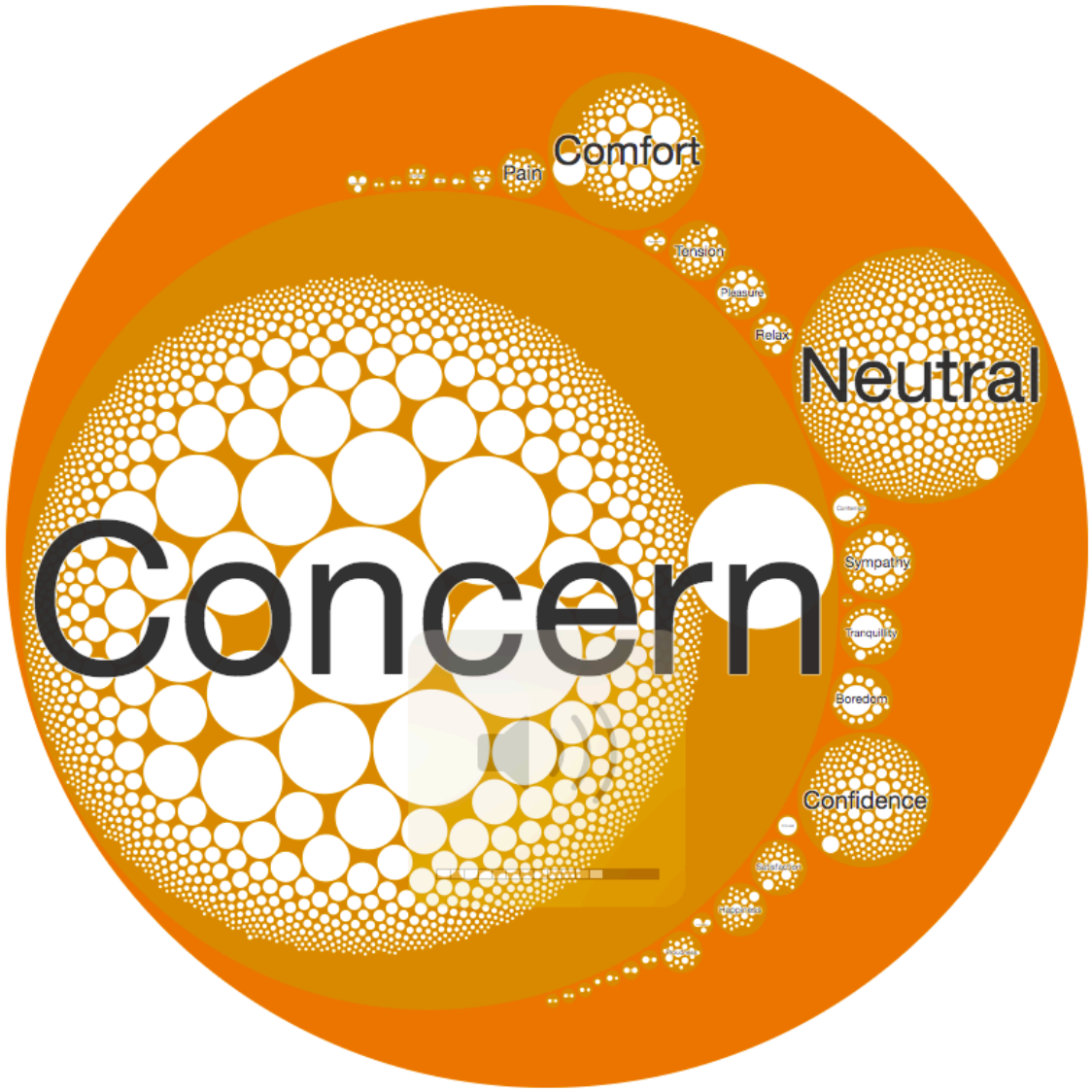


Graph 2-6 The topic Cloud based on emotional analysis enables the emergence of insights on dominant topics of discussions.

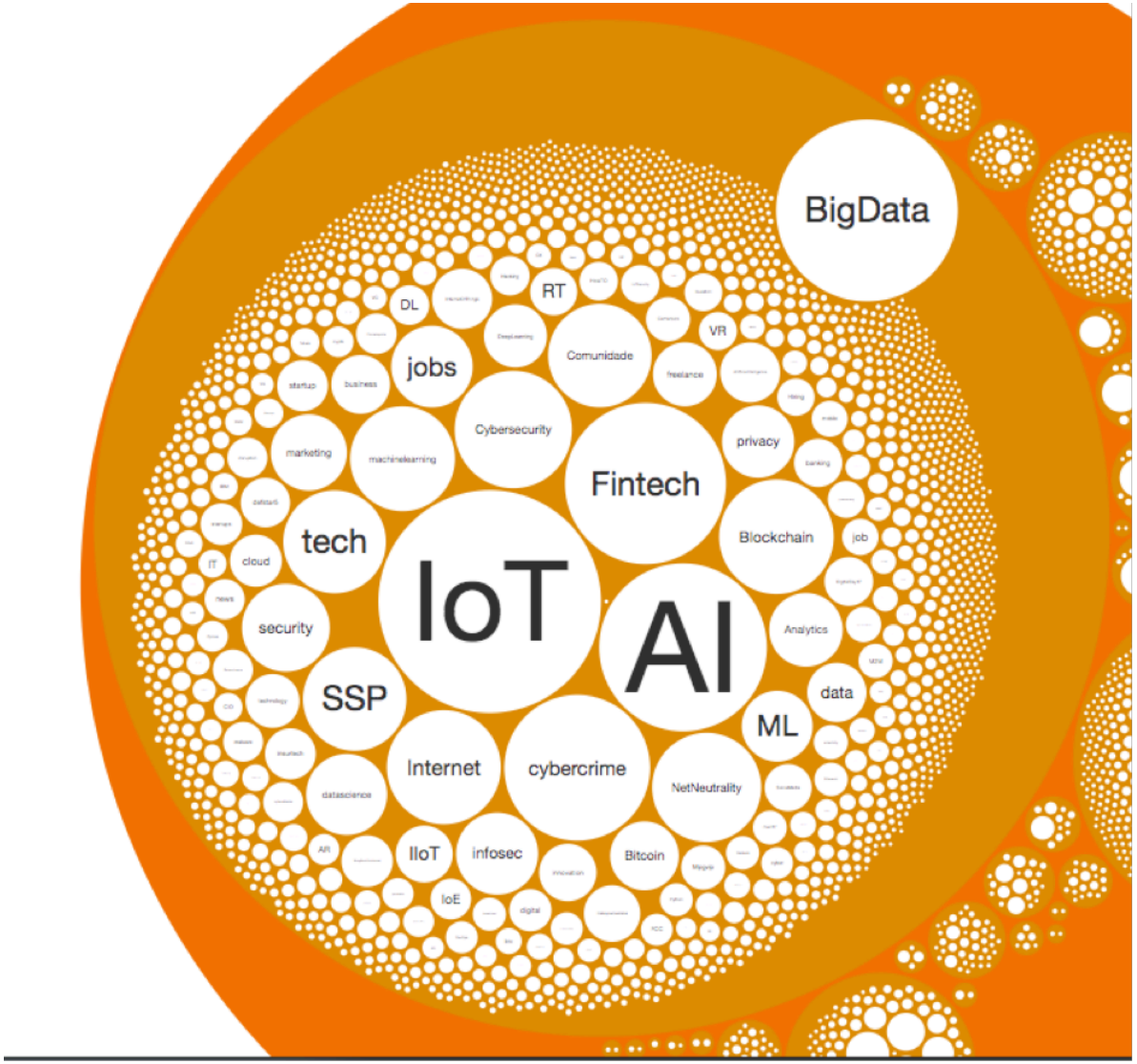
Graph 2-7 Overall Emotional Distribution for the subjects according to the Circumplex Model of Affect.

Graph 2-8 Overall Emotional Distribution for the subjects according to the Circumplex Model of Affect: Zoom in around the Origin.

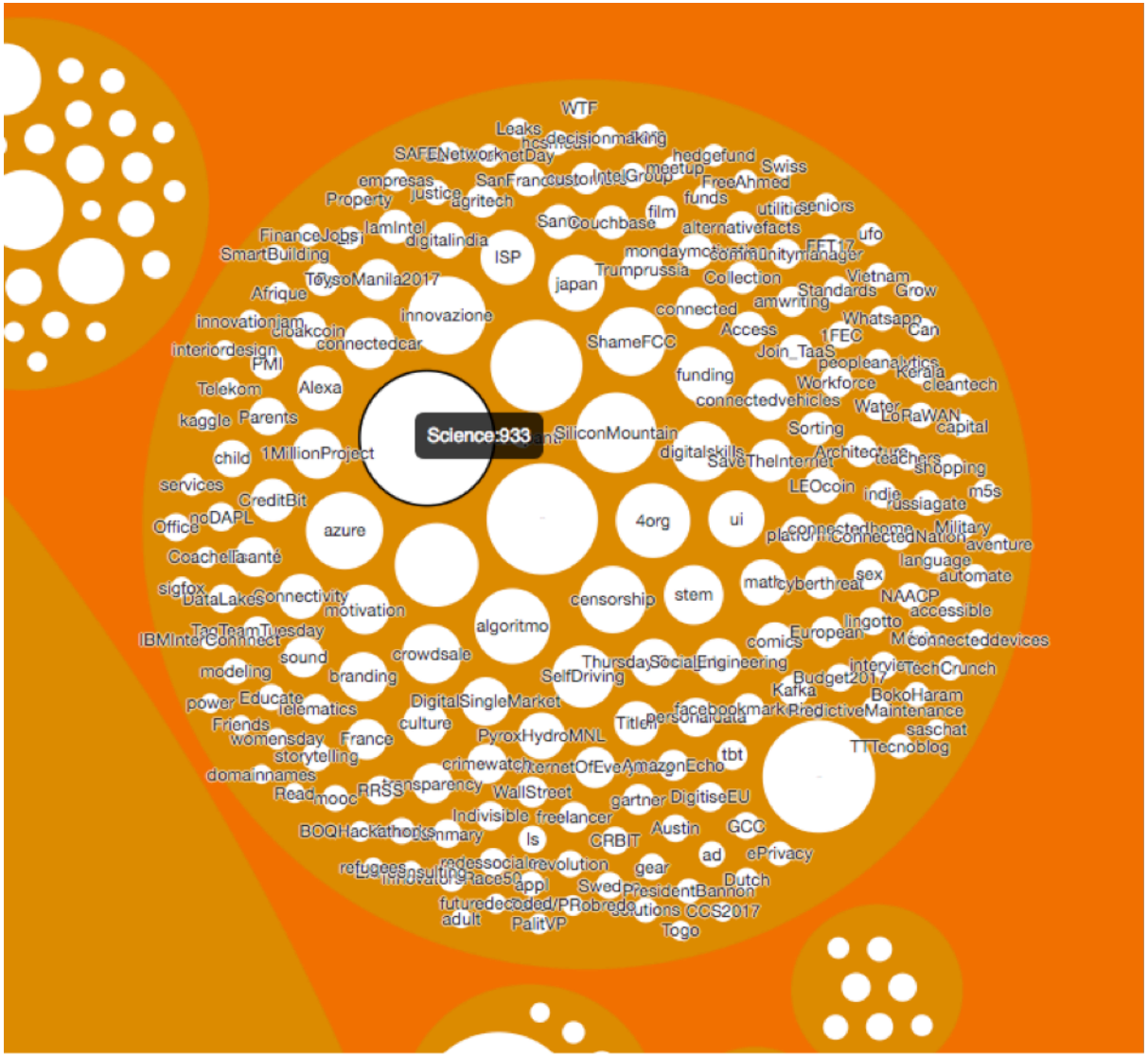




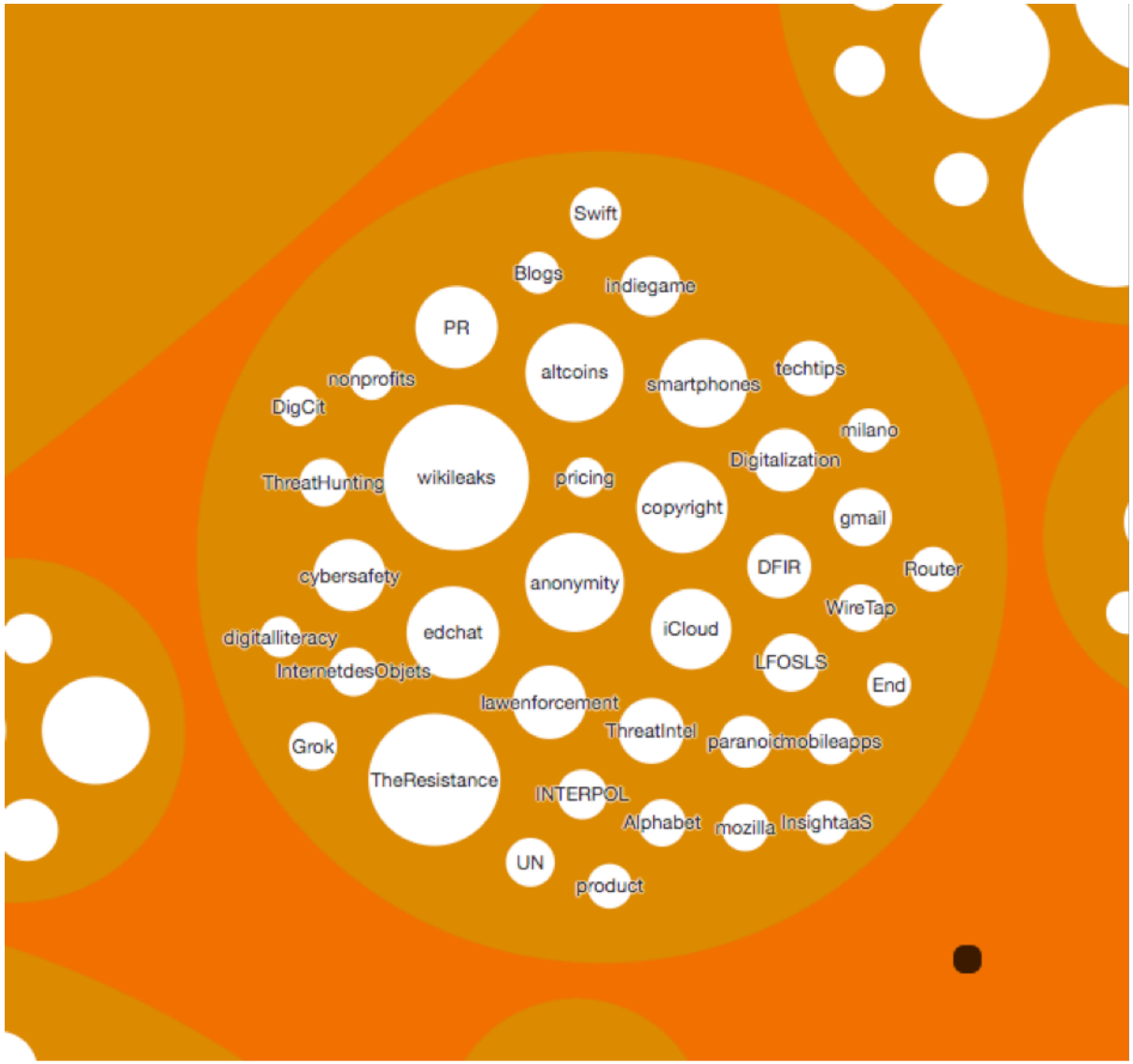
Graph 2-9 Distribution of Average Emotional Expressions by topic.



Graph 2-10 Distribution of Average Emotional Expressions by topic: Zoom-in on CONCERN emotions.



Graph 2-11 Distribution of Average Emotional Expressions by topic: Zoom in on CONFIDENCE emotions.



Graph 2-12 Distribution of Average Emotional Expressions by topic: Zoom in on TENSION emotions.

However, the scenario changes if we group the emotions expressed for the single topics, and we average them.

In Graph 2–9 emotional expressions for each topic have been summed-up and, for each topic, only the prevalent, most abundant type of emotional expression has been shown.

The visualization shows, in each bubble, the topics which present the corresponding emotion as their main emotional expression. Within the larger bubbles, each white bubble represents a topic, and the size is proportional to the number of emotional expressions of that kind for that topic[5].

Indeed, as showed in Graph 2–10, most of the topics fall in the “Concern” emotional state, meaning that they are in the slightly-negative-comfort and slightly-positive-energy quadrant. This corresponds to civilised discussions around problematic issues, where there is a high degree of uncertainty concerning possible angles and approaches to be taken. People engaging in this kind of conversations are not merely providing information – as in the “Neutral” cluster -, but are actively seeking for opinions and knowledge, posing questions and highlighting doubts to be discussed with others.

Also worth noticing and explaining is the "Confidence" emotion, in its topic-clustered version shown in Graph 2–11. While "Confidence" is a moderately positive emotion, some very negatively characterized topics appear within it, such as Censorship. This is because the Confidence emotion, according to our circumplex model of affect, is a mildly energetic, highly comfortable expression, usually corresponding to situations of certainty. This is why negative topics can appear within it: close inspection of the contents reveals how, for example, subjects are certain that censorship will be a (certainly negative) feature of the future internet.

Other interesting topics which appear in the "Confidence" emotional expression are connectivity, culture, multiple forms of artistic creativity (for example comics), access, justice, science, multiple topics related to "startups" and their ecosystems, and some typically “European” topics such as EU60 and Digital Single Market, for which Confidence is paired with very positive attitudes, denoting trust and certainty about a better future.

Finally, it is interesting to take a closer look to the TENSION cluster of emotions (Graph 2–12), which include topics around which a very active but not strongly polarised debate develops, with people showing some level of discomfort and uncertainty around the issues debated. Interestingly, highly controversial topics such as Wikileaks, The Resistance, anonymity and Altcoins but also references to organisations such as the UN, Interpol, Alphabet and nonprofits or to processes like “digitisation” and law enforcement.

[5] The exact numbers for these graphs are available in the Open Data released with the report (see the Open Data section).

2.3. Profile segmentation

We explored whether meaningful segmentations could be carried out to highlight particular systematic patterns in expressions.

We took a selection of the most important topics (IoT, BigData, AI, Fintech, CyberCrime, BlockChain, CyberSecurity, NetNeutrality, MachineLearning, Privacy, Analytics, Marketing, DataScience, Business, Jobs, Security, Education) and we used neural network based Machine Learning techniques to see if there were significant recurring patterns: we used a classificator[6] to see if there were systematic correspondences in the ways in which people expressed about these topics.

We discovered **3 profiles**[7].

Each profile is described according to the typical emotional expressions of that profile in respect to those topics.

Each profile is also given an average age: it is the result of the calculations performed as shown in the methodology session to infer in probabilistic ways the ages of internet users.

You can see the three profiles in Graph 2–13 (with the number of users that have been detected to be at least 75% in that profile).

In the Graphs 2–14 to 20–16, you'll find each profile's emotional distribution across the selected topics.

The **Optimist** profile has a strong, positive opinion about all the hot topics related to the Future of the Internet. It follows the themes of Data, IoT, Security, Business and Education in very active ways, expressing enthusiastic opinions which leave very little space for critique. It wants something positive to happen and imagines it being directly connected to technological innovation.

The **Activist** profile creates frictions in discussions on the Future of the Internet, tending to highlight the dangers and negative impact potential of most technological innovations. It is a very useful profile from the point of view of catalysing discussions, since it challenges people's assumptions, causing them to take sides. It expresses critique which is smart and informed[8], but leaves little space to opportunity.

[6] A classificator is a software program which uses Machine Learning to understand if there are naturally emergent clustering in a certain data set. It is very useful, as it does not require any pre-existing knowledge about the data set and, thus, allows discovering new features about it.

[7] Each profile is not exclusive, meaning that a subject (a social network account) does not belong to one OR the other, but, rather, it belongs to all three according to a degree: for example, account X may be 92% profile 1, 24% profile 2 and 38% profile 3.

[8] Indicators for "smart", in this case, are derived from text-complexity, such as L2, Flesch-Kincaid, Gunning Fog, which show a measure of how complex the text is. As for the "informed", it deals with the links and references contained and cited. Links are not all the same and they are evaluated in multiple ways, using blacklists, WoT, pagerank etc. Thus, we can evaluate how well informed the posts are, judging from the quality of the sources.

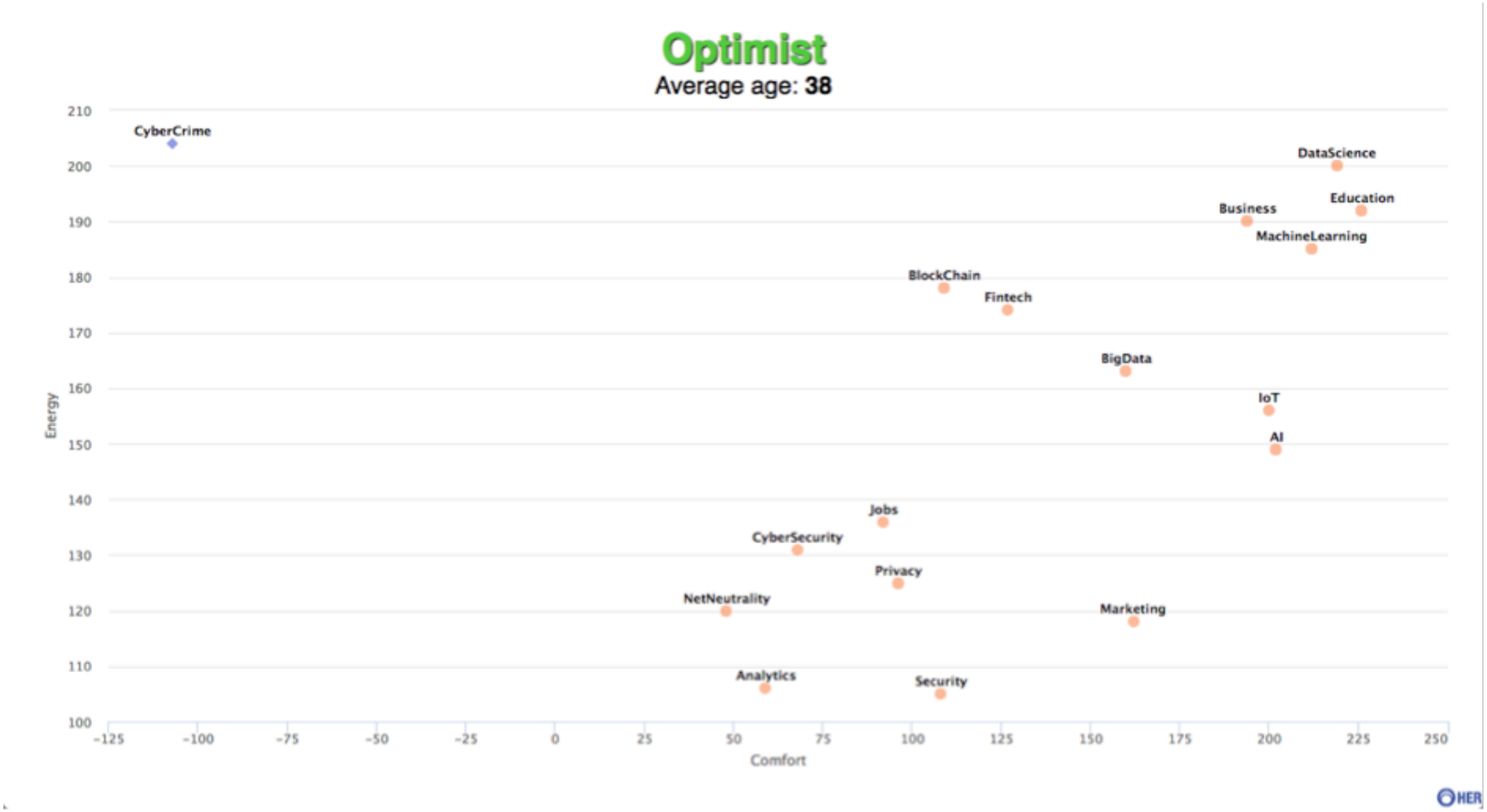
The **Exploiter** profile focuses on the business potential of next generation internet technologies. It does not really raise issues, and is only vaguely interested in the more critical topics, such as Net Neutrality or Security (unless they imply some sort of business opportunity). The profile is very interested and active in understanding how things can be applied to generate novel businesses. It is also actively engaged in the debate around education and skills needed to thrive in a fast-changing job-market.

The profile segmentation allows us to immediately visualize some interesting aspects of the ongoing debate around next generation internet technologies and their socio-economic impact potential. Indeed, even though the 3 profiles present very well-defined characteristics and attitudes, they all share a few hopes and concerns. AI; machine learning and the blockchain are unanimously seen as the most promising technologies in terms of positive impact-potential on our societies and economies, while, cybercrime is unanimously seen as a major threat. Fintech is a “high-energy”, much debated topic across the three profiles, but while the Optimist and the exploiter are very optimistic about its impact potential, the Activist is concerned about security, privacy and access. Interestingly, only the Activist profile seems really concerned (and vocal) about Net Neutrality, while Privacy – another very central topic – is seen as a major concern by the Activist, a major business opportunity by the Exploiter and a relatively uninteresting topic by the Optimist.

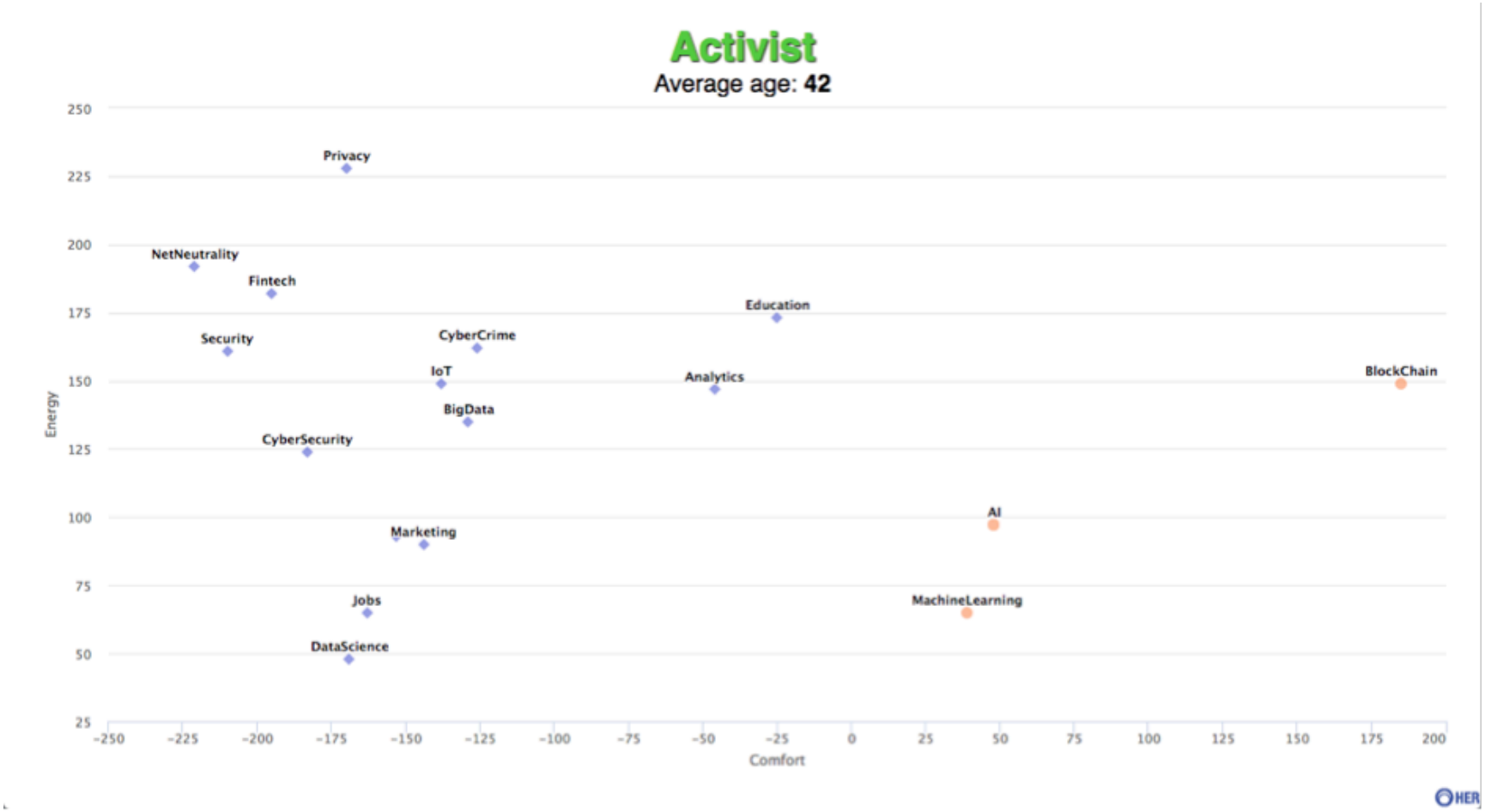
In this way, it is possible to compare how "smart" and "informed" messages for the various profiles are, and those for activists result more "smart" and "informed" than others.

Graph 2-13 Users talking about FoI can be classified as Activists, Optimists and Exploiters.

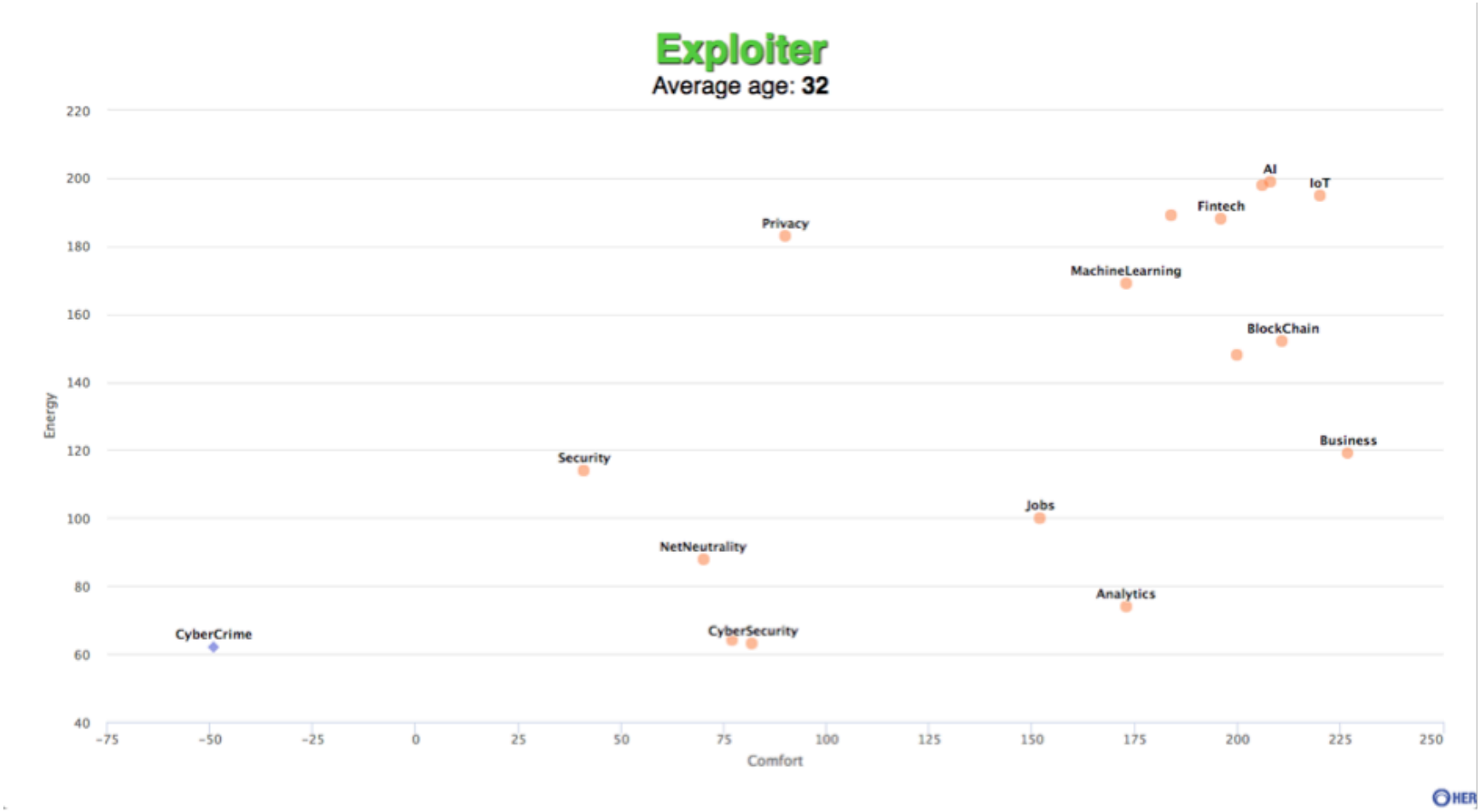
Graph 2-14 The Optimist Profile.



Graph 2-15 The Activist Profile.



Graph 2-16 The Exploiter Profile.



2.4. Topics Relations

Graph 2–17 is **a general map of the main relations among the topics** detected in conversations about the FoI.

Let's start analysing. There are **6 main clusters**, and multiple minor ones.

Purple: is the main technological cluster. It includes all the most important technologies which are deemed to be connected with the evolution of the Internet: IoT, BigData, Artificial Intelligence, Fintech, Machine Learning, Blockchain, Robotics. The cluster also includes many of the applications people think these technologies will have, from predicting events and behaviours to changing the way money and finance work.

Orange: this cluster includes conversations about the Cloud and Security, two topics perceived as closely connected. The discussions about Cloud and Security go hand in hand: they are sometimes very technical (and this can be seen by the number of acronyms which appear among the connected topics). They focus on the evolution of storage services and databases. Security (and privacy, and being able to determine access) is the primary concern. Other interesting connected topic: jobs. The cloud (and cyber-security to protect it) is very often discussed in terms of the job opportunities it may provide: for backend, frontend, maintenance, consulting, and new start-ups just to mention a few.

Light Green: the Future of Europe. This cluster is influenced by the occurrence, within the period of observation, of Europe's Digital Days 2017, on the occasion of the 60th anniversary of the Rome Treaty. The event generated a surge in conversations about the FoI, where the pillars of the EU Digital Agenda and Digital Single Market Strategy were clearly recognisable in form of key-words such as jobs, education, skills, industry 4.0, E-government, addressing inequality among others. Specific case studies and best practices are also mentioned, such as for instance Smart Clothing and Smart Wearables.

Gray: the Internet, and its main concern, Privacy. This cluster is focused on the ways in which users have identified Internet's evolution with problematic developments. Which does not mean "negative", but "problematic" in the sense that there are issues in this cluster which need to be addressed. The principal concern is Privacy, closely connected to Identity. Then come Finance and Economics, and the role of Banks and the relationship with Macroeconomics (for example through GDP). Then comes the issues of Education and learning.

Light blue: Cybersecurity and Cybercrime. This is a fundamental cluster: it is large, central in all discussions, and also connected in different modalities (one can notice a darker shade of blue-purple below the azure one, in which these topics are directly and very strongly connected to IoT, indicating the very high level of attention to the risks associated to IoT and Cybersecurity). This cluster has different impacts on privacy related issues, on businesses, on theft (also of the Identity kind), on all sorts of malicious software (an issue which is attracting increasing attention also from non-technical communities). It is also worthwhile noting how people express on these topics associating

them with legal issues (e.g. requesting more effective laws), and in relation with the global geo-political scenario.

Pink: Algorithms. Algorithms are dealt with in ways that oscillate from very positive, to negative or even scared and angry. On the positive side (which is more abundant than the negative one) algorithms are mentioned for their promising effects on medicine, in understanding "intelligence" and bringing about smarter systems, services and products. On the other hand, algorithms are connected with negative trends such as privacy violations, surveillance and control.

In **light grey**, on the top-left side of the diagram, is the discussion about Net Neutrality. It is a small cluster, and it is peripheral and unconnected with most other discussions and topics. Only a few people (yet sometimes very vocal) talk about Net Neutrality in the "classical" terms of, for instance, SOPA, PIPA, Broadband or Privacy.

Interestingly, it is far more common to speak about Net Neutrality with reference to operators such as Youtube and Netflix. In some cases, the existence of these operators and their technological architectures (such as geographical CDNs, Content Delivery Networks) are the subjects of strong critique; so for instance, in one post, the existence of operators which can manage their content through CDNs is interpreted in terms of "impossibility of Net Neutrality".

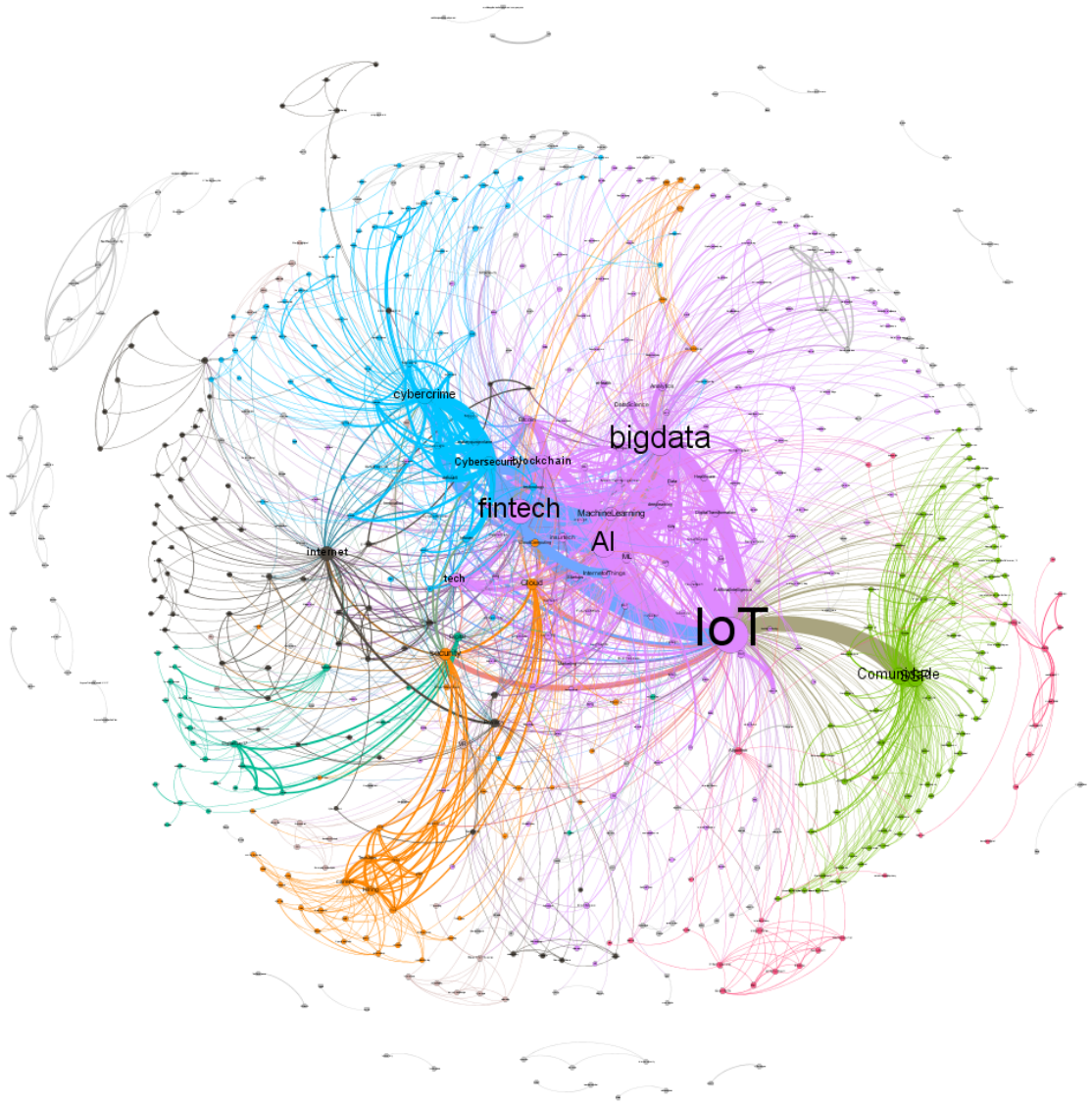
The green cluster on the left, revolving around the “Comunidade” key word, is an example of bots generated activity, with a multitude of messages generated automatically due to monitoring systems of different types (air quality, traffic, temperature, others) which constantly emit their recordings, tagged with hashtags which have been picked up by the software, as they are interesting for the research. These automatic contributions have been excluded from the considerations expressed in the following sections of the report, where you will find, a series of thematic focuses will detail the most interesting findings.

How to read a graph

To better highlight the principal connections, a threshold has been established both for the **importance** of the topics shown (at least 20 people must be talking about them) and for the **connections** (at least 5 people must have performed the association, otherwise it does not show):

- The larger circles show the most recurrent topics;
- The colours indicate clusters in which topics are systematically addressed together (ie: if two circles are of the same colour, it means that users systematically talk about them together).;
- The thickness of the arcs between circles indicate how strong is the connection (that is, proportional to the number of times they have been mentioned together).

You can read all the graphs using the same scheme.



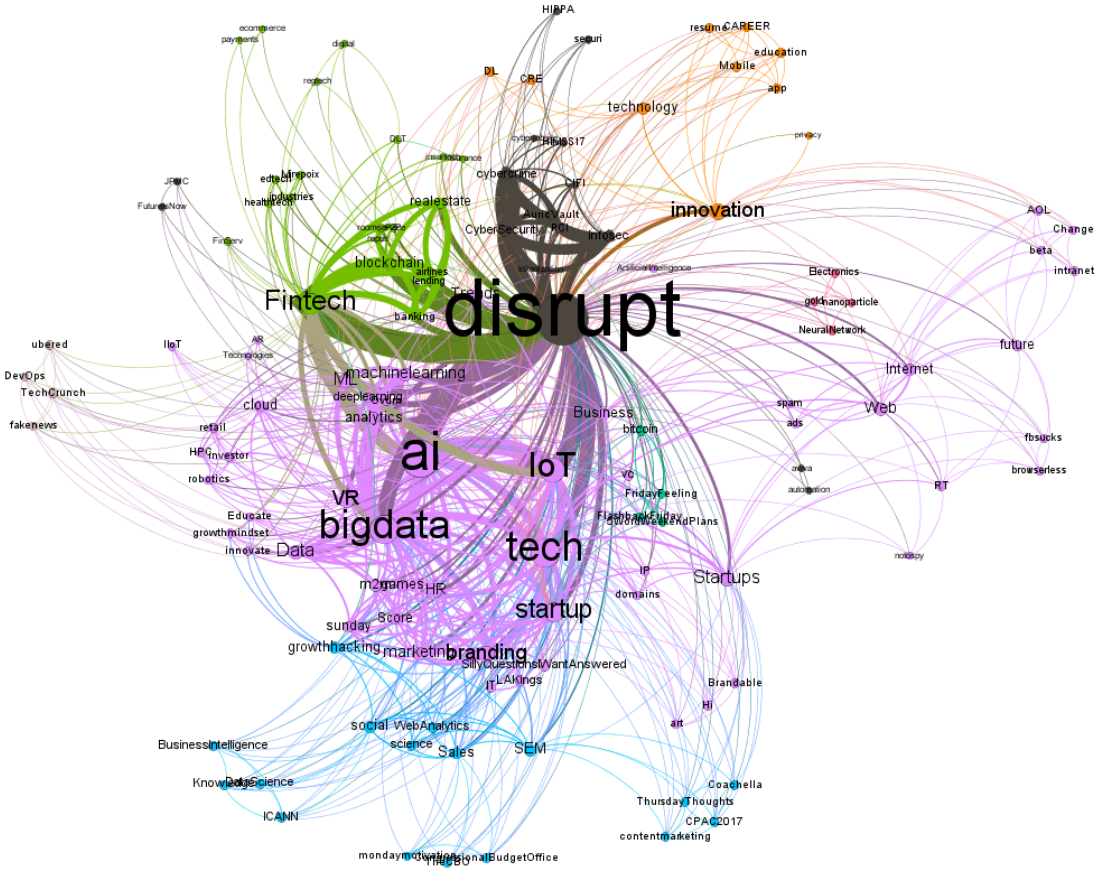
Graph 2-17 The conversation topics and their relations

3. Thematic Areas

3.1. Focus Area: coping with disruption

As already mentioned in the previous section, the first research question we wanted to answer through this network and sentiment analysis was about what emerging technologies would have more deeply changed the internet as we know it in the next decade, and what might be the impact of these technological advancements on our daily lives, as well as on our economies and societies. Before looking at impacts (in terms for instance of future jobs, but also in terms of problematic issues such as privacy, security, net-neutrality or democracy), let's have a look at the debate around "disruption" and about some of the most disruptive emerging technologies

3.1.1. Disruption



Graph 3-1 The relations of the topic “Disrupt”.

The clusters of concepts in this visualization clearly show which themes are perceived as "disruptive" in the near future: ***AI*, *Big Data*, *IoT*, *FinTech, VR and, to a lesser extent, blockchain*** *are by far the most disruptive technologies*. Moreover, their connection with the topics ***Startups*, Innovation, *Tech* and *Education***, suggests that their role is seen as increasingly pivotal in the business area.

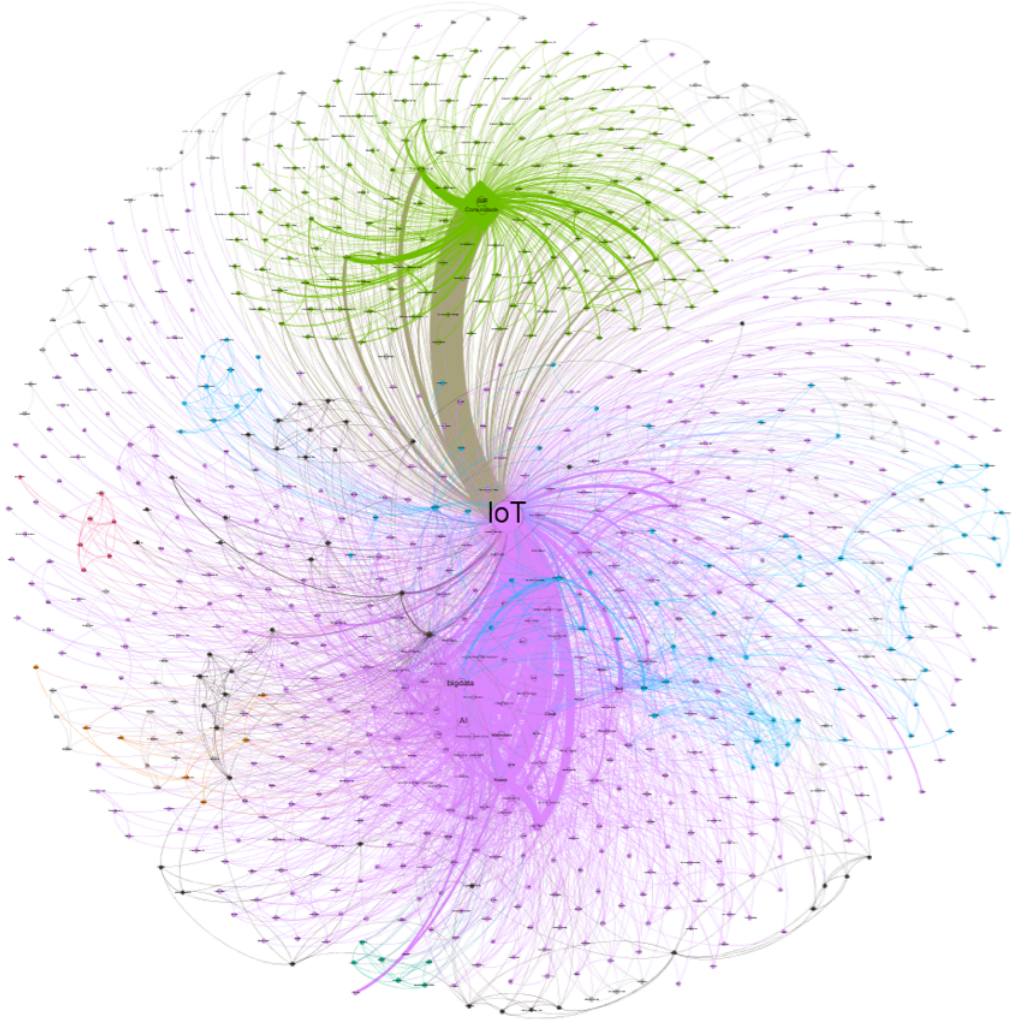
Digital *Disruption* is seen as having a strong connection with financial services, through ***FinTech***: what once was a buzz word is seen now as a fundamental layer of the digital disruption and transformation. The volume and connections of this topic suggest a great interest in this field, although its actual impact on the banking system is not clear. Users perceive that *FinTech* services will disrupt traditional ones: *Financial Intermediation* and *Risk Management* will be affected, as well as *electronic currencies*, *Ecommerce, Insurance* and *Online Payments*. We know that *FinTech* companies are already making big investments in *Blockchain* technologies and services, and this is evident also in the topic connection shown in the graph.

The ***Web Marketing*** topic is solidly connected to *Business Intelligence, Data Science*, and *Growth Hacking,* the new Marketing mixed approach developed in Silicon Valley. This is in turn connected to Social Media and Search Engine Marketing topic (*SEM*): Machine Learning algorithms and AI technologies could potentially make a clean sweep of the keyword-based Marketing, with an expected deep impact in the current Digital Marketing strategies (SEO, SERP, etc.). In sum, the future Web Marketing is seen as more and more based on Business Intelligence systems and Data Science knowledge and skills.

From the topic relations in the graph it is also possible to infer an emerging demand in the **organization of *Work***: the need to dramatically change current *Intranet Collaboration* practices and tools. This need for a profound evolution of this set of technologies and organizational practices emerges also in connection with the *Beta Tests* topic in the graph.

The *Fake News* issue emerges in the graph with a particular connection to the creation of effective removal tools, which are perceived as an actual demand from users.

3.1.2. IoT (Internet of Things)



Graph 3-2 The relations of the topic “IoT”

The **purple cluster** of this focus highlights the links with all the possible IoT applications. The most important ones are: **SmartCity**, **Robotics**, **Fintech** and **Health Care**. The huge impact of the IoT on the industrial sector is represented by the **Big Data**, **Manufacturing** and **Industry 4.0** topics, but also by the brand **Amazon** (supposedly thanks to its Echo device).

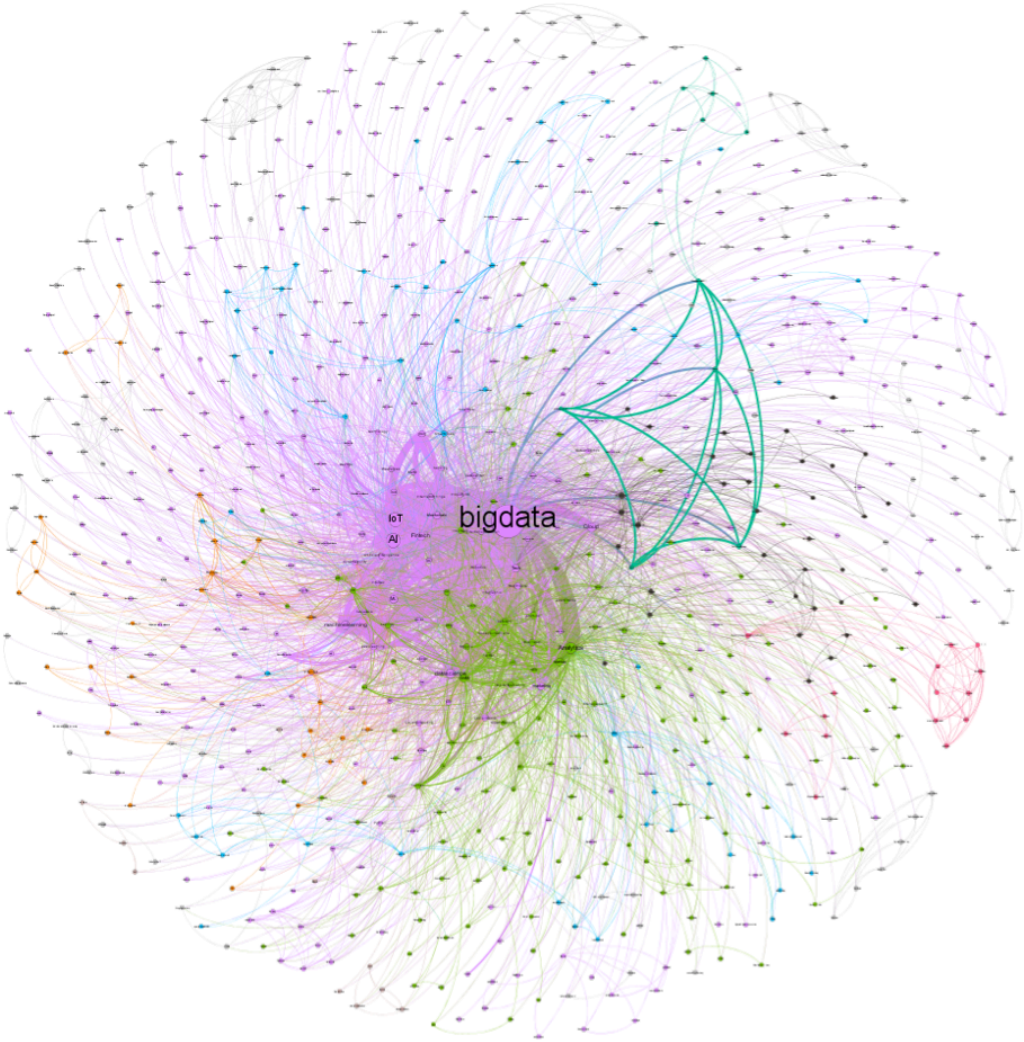
As described in other parts of the report, there are also multiple connections between IoT and topics such as privacy, cybersecurity, safety and security in general. This shows both the concerns expressed by operators and citizens (and reported in the news), and the business opportunities recognised in this space (for example with startups communicating about concepts in cybersecurity and privacy applied to IoT). There is also a degree of activist interest in Privacy and Security related issues on IoT.

The connections of this cluster highlight also the need and demand for IoT systems protection from cyberattacks, due to their well-known vulnerability, and for **Data Analysis** and **Data Science** systems and solutions, in order to interpret correctly and obtain meaningful insights from the huge amount of data generated by IoT systems.

Another interesting cluster is the **light blue** one, referring to new jobs and careers (see topics **Hiring, Career**, **Jobs**): within the IoT-impacted economic sectors, it is already registered a primary need for seeking and recruiting people with new skills and competencies. It's significant, in this context, the connection within the same cluster to the topic Learning, which in turn connects to the SmartHome solutions and systems subcluster (Microsoft OS, iOS, etc.).

The connection with the **Startup** ecosystems is present, as expected, but not significantly IoT-related and often noise biased.

3.1.3. Big Data



Graph 3-3 The relations of the topic “Big Data”

The **Big Data** topic is very correlated with **Analytics** and **Data Science**, as the **purple cluster** shows clearly. The topics linked in the cluster suggest a trend in the usage of Data Analysis tools for business development: in the centre, we note a whole set of topics revolving around **Analytics** and **Machine-Learning Algorithms**, **Neural Networks** for **Deep Learning, Artificial Intelligence** and increasingly more accurate predictive analysis systems.

From the point of view of the opportunities that are opened up by BigData, the topic is seen as a major **enabler of businesses** based on IoT, AI, Fintech, Blockchains, Wearables, Machine Intelligence, Sensors, Clouds, Smart Everythings and, in general, of all the most disruptive opportunities.

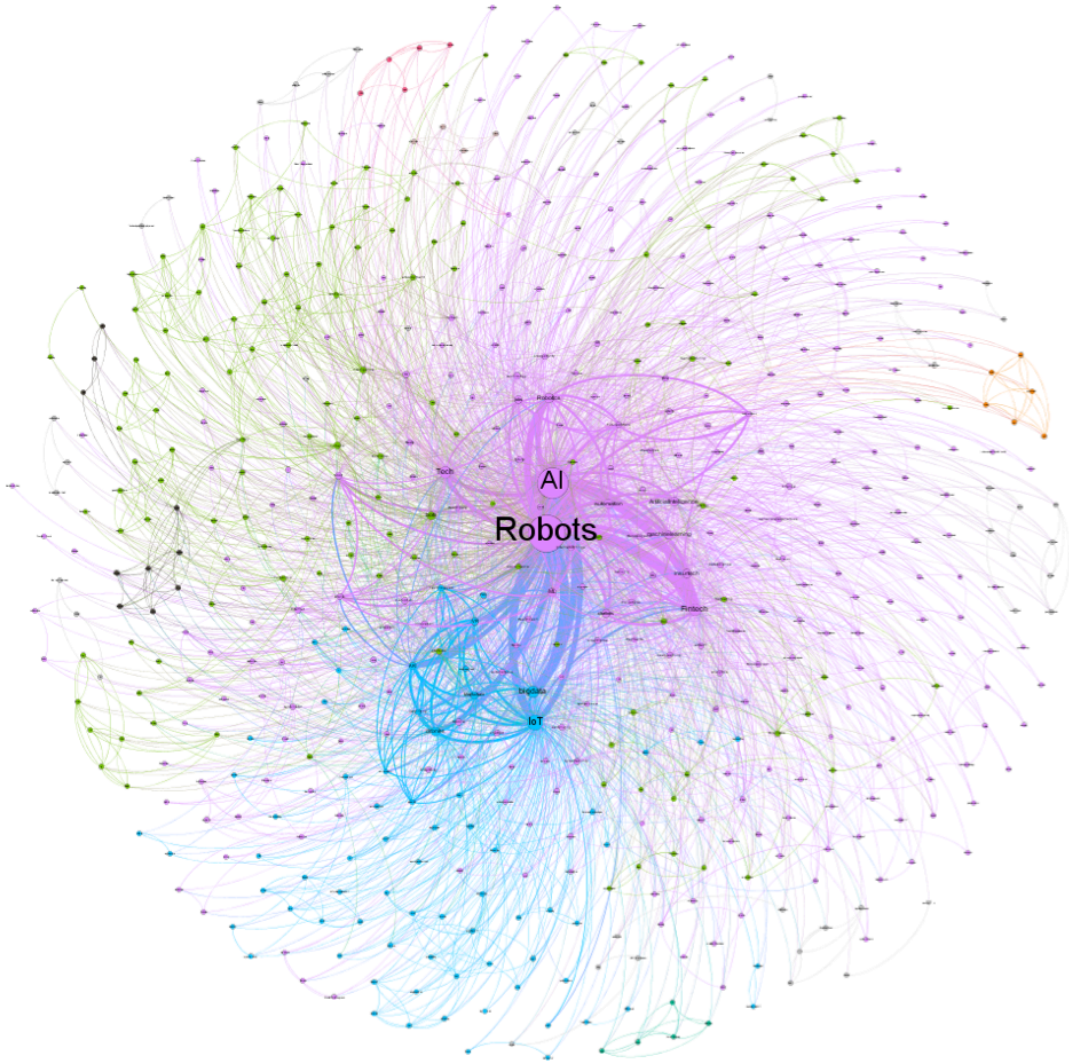
Another modality of the **Analytics** topic is found in the **green cluster**: it is connected to **Startup**, **Digital Marketing, SEO, Social Media, GrowthHacking, UX**, suggesting that the startup ecosystem is perceived as mainly developing around these technologies and solutions, mostly delivered **via SaaS modalities.** In this cluster we can also find the topic **Job,** connected on one hand to the need of finding people with Big Data analysis skills, and on the other hand suggesting a growing interest towards employees with an expertise in **Virtual Reality,** certainly a growing and trendy business sector.

Tools and competences needed to perform data extraction and analysis, e.g. **Linux**, **NoSQL**, **Apache Sparks**, **Java, etc.** are distributed between the **purple, the pink and the grey clusters.** The **UX** topic presence suggests a growing attention to the needs of “non-technical” user skills, typically somebody with business management competences. Here emerges the brand Amazon Web Services (**AWS**), which is considered to be a standing example of how to effectively offer Big Data related services in ways which are accessible and comprehensible to management.

Another big topic is the one related to **Big Data** and **Agritech** (**blue cluster**), as data analysis can improve substantially agricultural production, for example through Precision Agriculture.

Highlights: the evident **aqua green cluster** is about the recent French Presidential elections and the digital analysis performed during the campaign to predict their outcomes.

3.1.4. Robots and Artificial Intelligence (AI)



Graph 3-4 The relations of the topic “Robots”

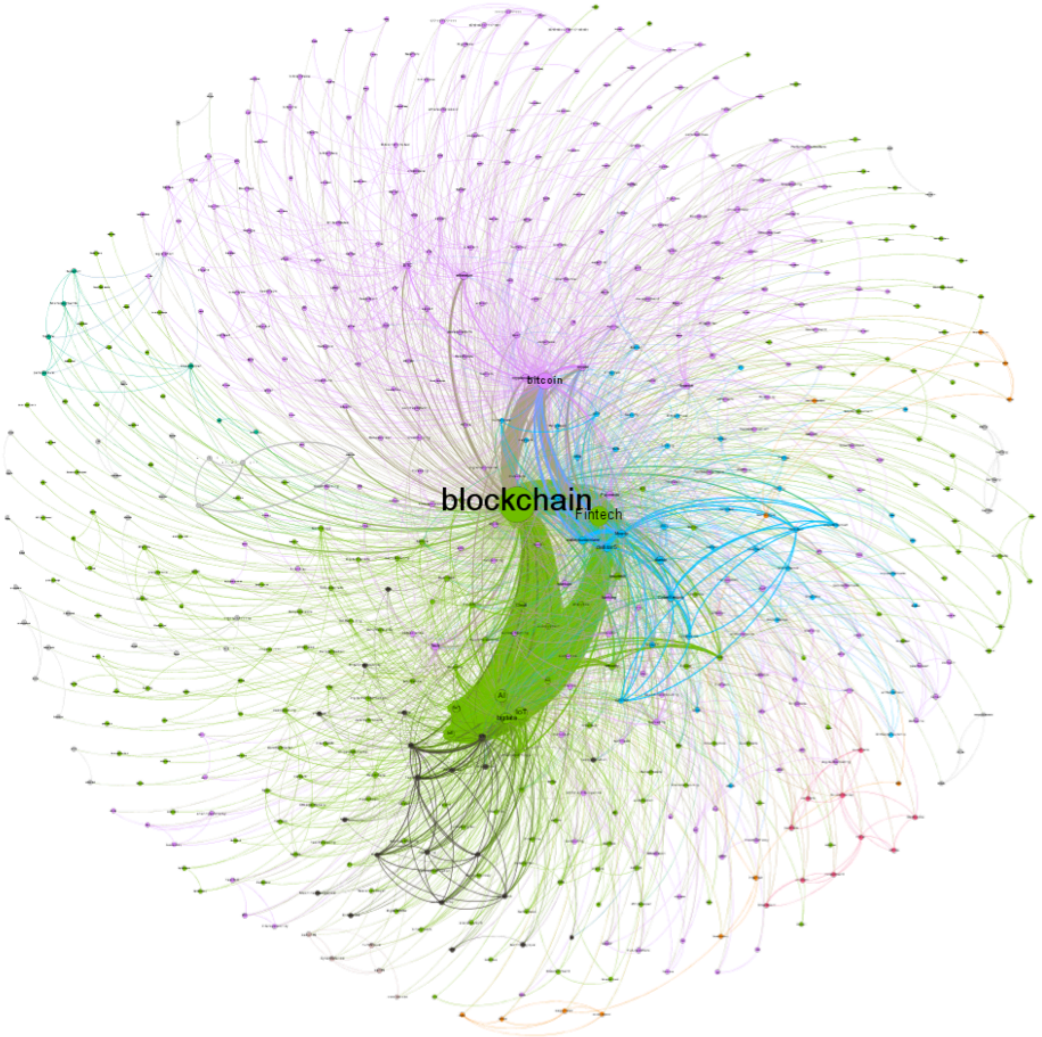
The topics **Robot** and **Artificial Intelligence** (**purple cluster**) and **IoT** systems (**light blue cluster**) are very close and have almost the same weight: this means that **they are highly connected and correlated to each another.** Moreover, they are evolving independently and continuously, intersecting together in a quasi-symbiotic way.

Within the **purple cluster** various interesting links emerge. One of the strongest connections is between **AI** and **Fintech**: apparently the Bank & Finance industry is very interested in AI technology, either for investment purposes, or as a means for business and process innovations, e.g. to automise and optimise payments and transactions.

Another interesting link within the purple cluster is the one between **Robotics, Automation** and **Jobs, suggesting** growing attention for **business models cantered on robots, rather than humans**: in fact, the primary liaisons registered are between topics like **Future of Work** and **Jobs *for* Robots**. However, the debate on it does not represent a fundamental issue in the Future of Work debate, as we can infer from the weak connections amongst the **Robots** and **Human** topics. Moreover, we highlight a weak, but significant sub-cluster focussed on the **political discussion on Robots**, **Human-Machine** relationships and their impact on **Work** and **Society**. Two aspects emerge, strongly linked to the words **Future** and **Human**: the increasingly popular perception and expression that Robots will spread and relate with human beings, and the rise of investigations around future politics and regulation architectures (as indicated for example by the “future politicians”, “algorithms” and “robot politicians” topics). If in the next future Robots will evolve and spread, becoming indispensable, it is important to understand the consequences of this fact on Society and to initiate a profound political debate.

The **Chatbots** and **Bot** topics emerge from two different clusters. The first one stands out from the broader discussion of the purple cluster, deriving directly from the Fintech topic (see above), and particularly from the **RoboAdvisor** topic, this being one of the most buzzed innovations in this industry. The topic around **Bots** is instead to be found in the **green cluster**, revolving around **Social Media** and technological innovations in business processes and operations. What we can see in this cluster is that corporations are investing in increasingly smart chatbots and integrating them in customer service functions as well as in management systems and tools, such as **CRMs**. **Social Media chat platforms** such as **Facebook Messenger** or **Telegram** are making their use easier, and web-based services simplify their personalization and development by users who don’t have coding skills.

3.1.5. Blockchain



Graph 3-5 The relations of the topic “Blockchain”

Blockchain has rapidly become a trendy topic on social media, and one in which distinct types of subjects put their trust, highlighting the potential of this technology for innovation and disruption. From the graph, we can see first of all, in the purple cluster, **Blockchain's original use in Fintech**, **with Bitcoins, cryptocurrencies and related technologies**. The frequent connection in the **green cluster** with the topics **Banking**, **Security**, **Fintech** and **Finance** suggests that there is a strong demand for the creation of **Blockchains dedicated to the financial sector, with a wide use of AI and Machine-Learning** systems, in order to develop **new banking services**, with no more intermediation needed. In the same cluster, the connection with IoT, suggests how multiple types of solutions and technologies could come together, to enact complex architectures in distributed ways, going beyond, in subjects' expressions, current cloud or centralized schemes.

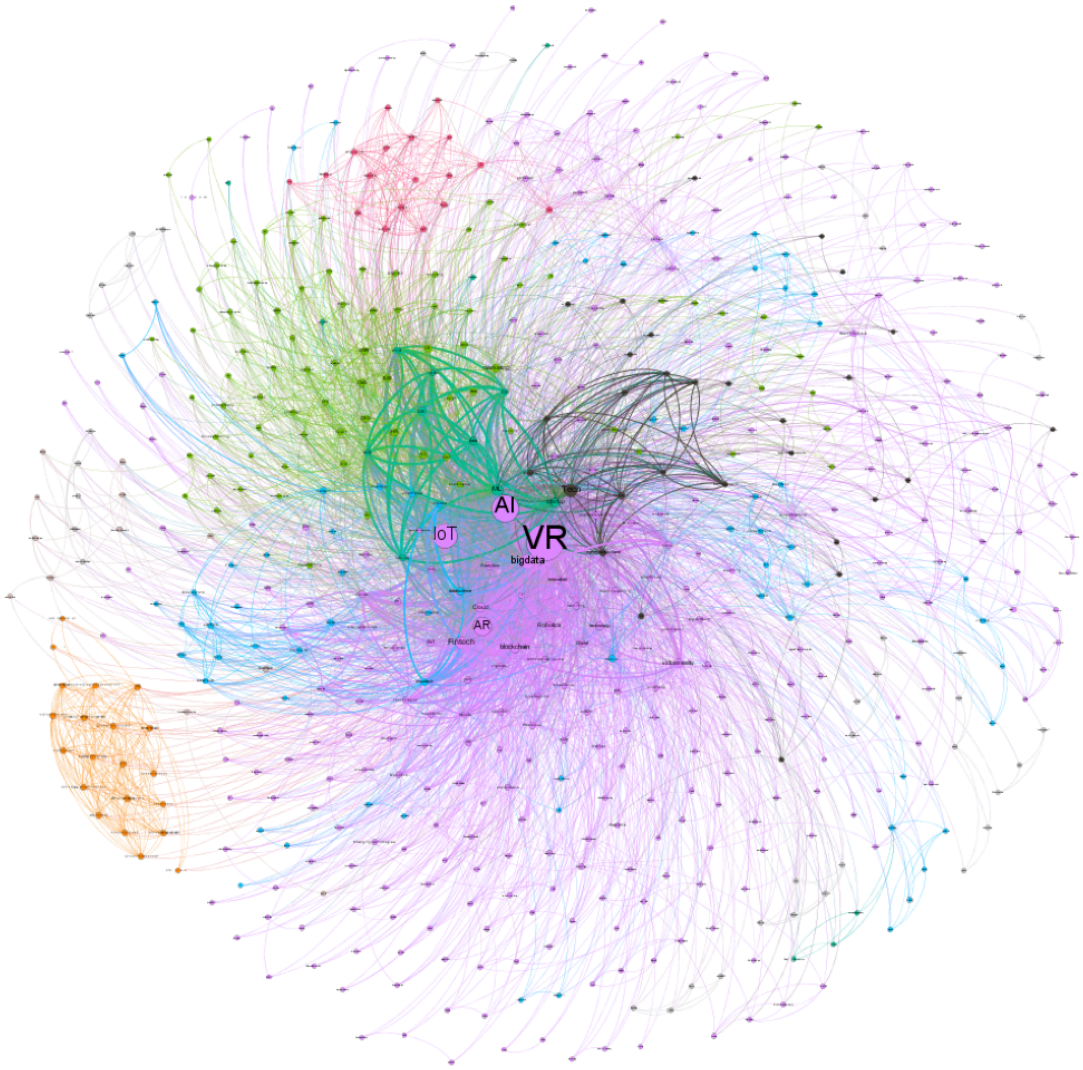
Secondly, there is a strong association with other technologies such as **AI, BigData, IoT, Robotics,** mainly because of their perceived disruptive potential on the market and on our societies. Conversations also show how VR is gaining some attention in relation to Blockchains, for example with the possibility of creating distributed virtual worlds, as shown in the lower-centered light grey cluster.

Thirdly, we can found a series of often very heated discussions in favour or against this distributed layer, with detractors normally focussing on performance issues (low speed, scarce scalability, high levels of energy consumption, etc.)

In the light blue cluster it is highlighted the CyberSecurity Startups ecosystems developing blockchain systems against CyberCrime attacks, links between blockchain and the so called "Dark Net" are also strong.

Curiously, in the **purple cluster we can find a lot of discussions** around the **P2P Blockchain platform Ethereum**, used as a cryptocurrency and contract transfer platform. The weight could also be due to the communications around its April Meetup.

3.1.6. Virtual Reality (VR)



Graph 3-6 The relations of the topic “VR”

Virtual Reality is a rich topic which, from subjects' expressions, could be expected to play a major role in the near future.

It is connected, in the **Purple Cluster,** with the technologies which, as shown in other parts of the report, have been more closely associated to disruption: AI, IoT, BigData, AR, Fintech, Robots and more. This denotes how radically new modalities for interactions with these technologies are expressed as an interesting and valuable area for research and innovation, potentially bearing strong returns on investment and positioning in disruptive markets.

In the **light blue cluster**, VR is connected to a series of specific technologies, which have explicit connections with data, the possibility for prediction and risk modelling, visualization and management, such as Insurtech, DataScience, MedTech, and Machine Learning scenarios.

VR is one of the most promising areas for education, with multiple topics in the **green cluster** dedicated at expressing the interest for Human Resources and Education, and highlighting how multiple practices (from photography to UI design, to advertising, and more) would radically change with VR.

VR's connection to AR has strong connections with **Android or iOS Mobile** services and games such as **PokemonGo**, or other **Indie Games**. The Gaming scenario constitutes a clear connection with education and training, specifically through the comprehensive list of skills and competences related to VR/AR, for example code languages needed for their implementation: in **bright pink**: **Java**, **Ruby**, **Nodejs**, etc. The explicit mention of **Job**, suggests and confirms relevance for the job market The new Jobs originated by VR developments are also present in the **grey cluster**, pivoted on **Startup** and **Smart Devices**.

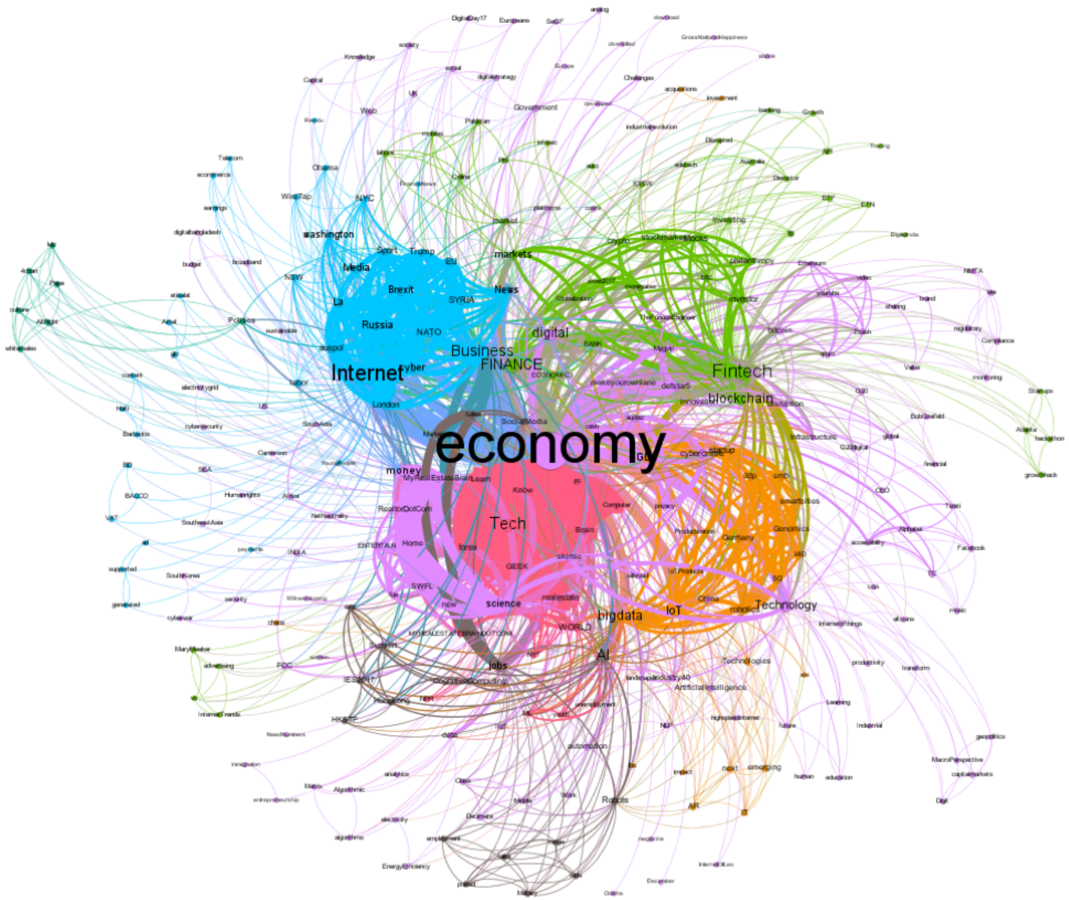
Going back to the **purple cluster**, we highlight the link between **VR** and **Fintech**: a potential future trend in the Bank and Financial sector. The applications can be varied, even though they are still in embryonic state: from geo referenced support in finding financial services nearby, to immersive visualization and interaction with data and personal financial advisors, supporting the investment decision making process. Looking at the **light blue cluster**, VR is also linked to InsurTech, enableing radical transformations in interaction paradigms, for example in CRM, with VR **ChatBots** replacing ordinary interactions.

Another interesting connection (**green graph**) appears between **Tech** and **Marketing,** suggesting that the development of Virtual Reality solutions for marketing purposes is potentially disruptive for current **Customer Experiences**.

3.2. The Future of Work

As emerging from previous paragraphs, conversations about the future of the internet are very often conversations about future employment and business opportunities. In the following sections we'll take a closer look at networks developing around work related topics, and, more in general about the economic potential of emerging internet technologies.

3.2.1. Economy



Graph 3-7 The relations of the topic “Economy”

Graph 3–7 shows the main modalities in which the word **Economy** appears in different conversations.

The **purple cluste**r is the one with most branches and connections. It offers a fundamental vision: **Economy is Digital, and tightly connected to Technology, Money, Fintech, Science, Data, Social networks and other types of platforms**.

Many discussions derive from communication initiatives and policy suggestions led by the informal G20 task force on Digitalisation. These discussions are preparatory to the next G20 Summit and aim at creating a multinational consensus around global policies in support of Digitalisation. The Focal Areas of this task force are: **Global Connectivity**, **Artificial Intelligence** and **Industry 4.0**. These topics are all very present in the cluster, as well as the following related ones: **Internet Access and Inclusion**, **Algorithm Accountability**, **G20Consumers**, **IoT**, **Internet Culture**, **Cybercrime**, **Bank Blockchain Consortium**. It's interesting to pinpoint the link between **G20Digital** and the brands of big corporations, such as **Facebook** and **Alphabet**. This derives from two different modalities of discussion: one includes **users' concerns for the protection of their data** when accessing online services. The other one emerges from Internet Companies' concerns about **US protectionism**, that could **potentially harm their multinational businesses.** Alphabet, Facebook and Apple aare among the most cited companies. Moreover, within the same cluster, we can find interesting connections between the words **Future**, **Artificial Intelligence**, **Automation**, **Industrial, Human**, **Work**: clearly, people are concerned about the impact of technological innovations on their lives, in terms of jobs of course, but also in terms of a much needed reflection around the role of politics and the public sector.

The **dark grey** relations highlight the discussion on the impact of **AI and robots** on the current work/economic relations among specialists and researchers: a discussion typically driven by media outlets (most of the conversations are news sharing).

The **light blue** word relation subset describes a complex map in which topics such as **"Internet" and "News" interweave with the hottest and more recent international geo-political phenomena**: from Syria to Russia, from Brexit to the American elections. The Syrian situation, for example, arises in a discussion on how p2p, decentralized models (such as blockchains or bitcoins) can be used (and are used) to fight ISIS and terrorism (eg.: "*#fintech Headlines! How Amir Taaki Tried to Build Bitcoin Economy in Syria While Fighting ISIS http://ow.ly/bHjl50ax8To*"). Or as in those threads in which Russia is connected to the fake news topic and the manipulation of Trump's presidential election results (eg: *#Russia #USA #WallSt #NATO #EU #Putin news #London #Trump #finance #Brexit #Macron #France #Comey.fired #Schumer .Hillary emails Russians. pic.twitter.com/PiA0bRVjPF*"). In general, **economic dimensions are matched to technologies** (especially disruptive ones, like blockchains) **and technologically-driven phenomena** (such as Fake News) to indicate possible or existing game-changing shifts.

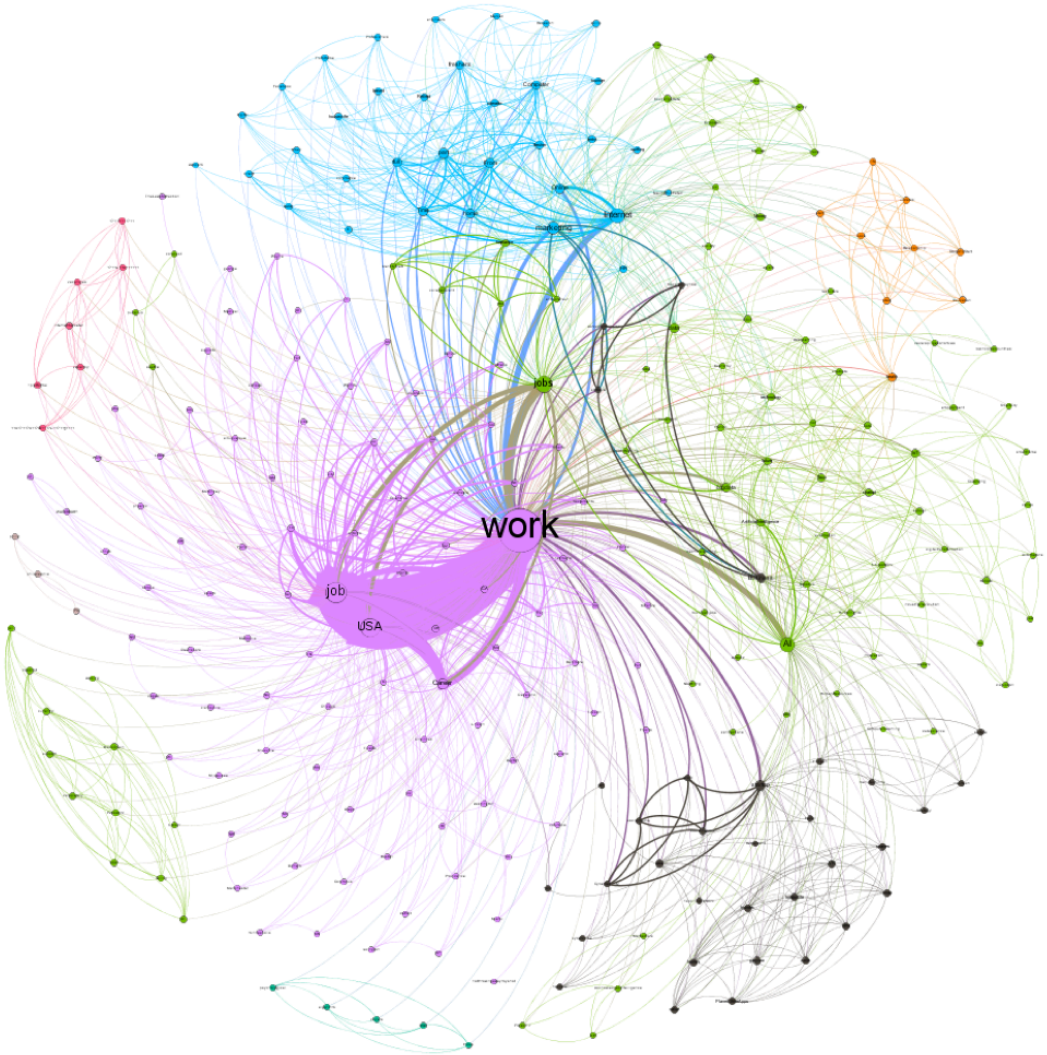
The **green graph** shows the connections with the term **Fintech**: from **Crypto Currencies** to **Investments** and the main changes affecting **Banks** and the **Financial** sector. Once again, Fintech is expected to impact Financial

Markets, transforming investments, and introducing novel financial flows, for example in peer-to-peer and distributed modalities, including blockchains and alternative currencies.

Expressions on this theme are excited and active, but they do not lack a certain concern: if on the one hand there are a multiplicity of subjects discussing about disruption and growth, on the other hand, as noted in the introduction, these are "apprehensive" expressions, meaning that they are not necessarily negative, but pose problematic questions that do not always find answers.

In **dark green**, in the upper left corner of the graph, is an unexpected and direct **connection between Economy** and the so-called **Meme Economy**. This new trend is related to socio-political (see for example the reference to the renowned meme "Pepe the Frog" and the "alt right" movement) and socio-economic phenomena (see the 4chan meme board). MemeEconomy is a typical web phenomenon, born from a 4chan sub-reddit, in which thousands of users are publishing the Memes trend valuations and forecasts, as if they were stock exchange brokers or financial investors. The MemeEconomy has spread in ways which recall financial-like indexes, such as the NASDANQ, the corresponding NASDAQ for Memes, and magazines, such as the monthly meme insider (see also this post and this one).

3.2.2. Work and jobs



Graph 3-8 The relations of the topic “Work & Jobs”

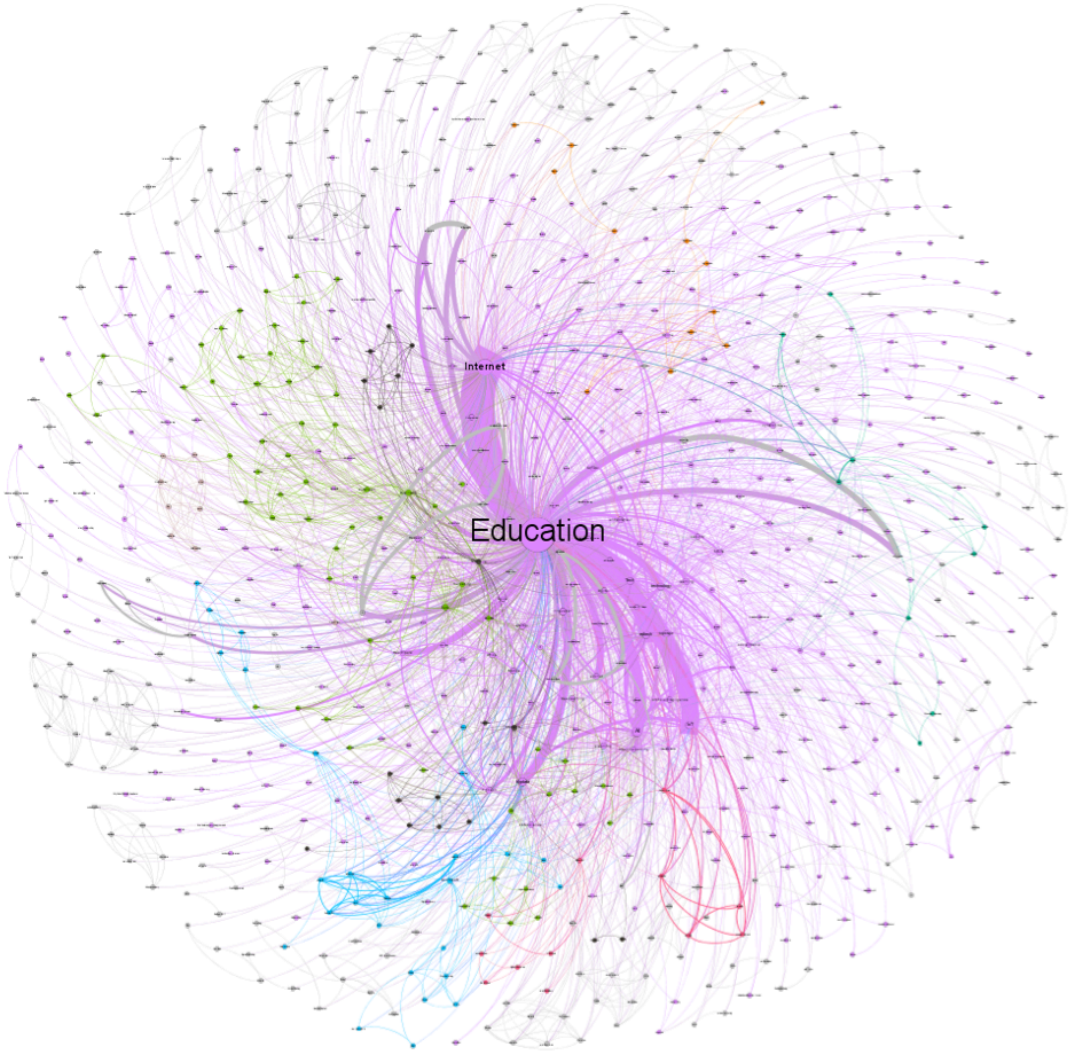
What jobs will we have in the future? The graph suggests that **people are looking at the *US Job* market to have an idea about new trends.** ,

The **development of *Bot* and *Chatbot*** tools, especially via integration with evolving *AI* technologies, is considered to have a great impact on certain specific jobs, e.g. on customer care functions. There are **two aspects** to this. On the one hand there is a heated discussion revolving around modalities in which **Smart non-human Agents may replace humans to carry out a broad range of jobs**, leading to a rise in unemployment. On the other hand, AI, Bots and other forms of **Smart Agents** are seen as a **great opportunity,** also in terms of creating new jobs. This is particularly evident in the education sector, and in several business scenarios, with Bot and AI design and development interpreted as key skills for the future.

Marketing is another recurrent topic, that has peculiar interconnections with other topics. Its links to *Home*, *Online*, *Part-time*, *Housewife* and *Freelance*, for instance, suggest **a significant boost towards *Agile* and *Flexible* work modalities**, together with an increasing distance from the classical "*worker's lifetime company loyalty*" model.

This is confirmed by the ***Startup*** node of the graph, assuming that a startup is a company iwhere innovative organizational paradigms are widely adopted. **Startups** are mentioned according to two main modalities. Firstly, as **stereotypes of the 2.0 work modality**: a large part of the subjects whose expressions were observed, expect that the most innovative jobs in the internet technology area will take place in Startups. Secondly, Startups are mentioned in relation to novel **market areas which appear particularly promising**, such as Cyber Security and Social Media.

3.2.3. Education



Graph 3-9 The relations of the topic “Education”

The discussion around the topic **Education** is wide and well defined: a large number of people and organizations are talking about it, initiating and fostering a large number of discussions on a large set of related topics.

The purple cluster sees **Education** linked to a wide variety of topics which constitute its principal associations: firstly, a large set of technical skills such as **Internet**, **BigData**, **Artificial Intelligence** (in many word modalities and languages, e.g. in french), **IoT**, **Science**, **Cybercrime**, **Fintech**, **Cloud**, **Machine Learning**, **Deep Learning**, **Data Science**, **Robots**, **Augmented Reality**, **Virtual Reality, Growth Hacking**, **Computer Science**, **STEM** (**Science, Technology, Engineering and Mathematics);** secondly **with transversal skills such as Entrepreneurship and Innovation, and thirdly in connection with specific sectors (such as Healthcare) or organizations (such as Startups** and **Government**).

These topics are used in a wide variety of contexts. Most notable are expressions of necessity, or the ones which define a lack or deficiency. For many subjects **Education should deal with all of the above technologies**, whether in school, or University, or subsequent professional training. On the other hand, subjects express how **citizens should participate in education processes**, including entrepreneurial skills, but also inclusion and social innovation.

Moreover, there is an interesting and explicit demand for Social Media competences in people's educational curriculum, as seen in the connection between the topics **Learning Social Media** and **Future Work**.

In terms of educational institutions and tools, together with College and Universities we find **Online** and **Mobile** courses and **MOOCs**. Another interesting note is the evident bridge role of **EdTech**, a technology-focused magazine for IT professionals teaching at K-12 schools and institutes of higher learning.

Very interesting is also the subcluster around the topic **Women**, presumably boosted by the discussions originated during the **InternationalWomenDay** in which the **WageGap** issue is discussed as well as the need to **BeBoldforChange**.

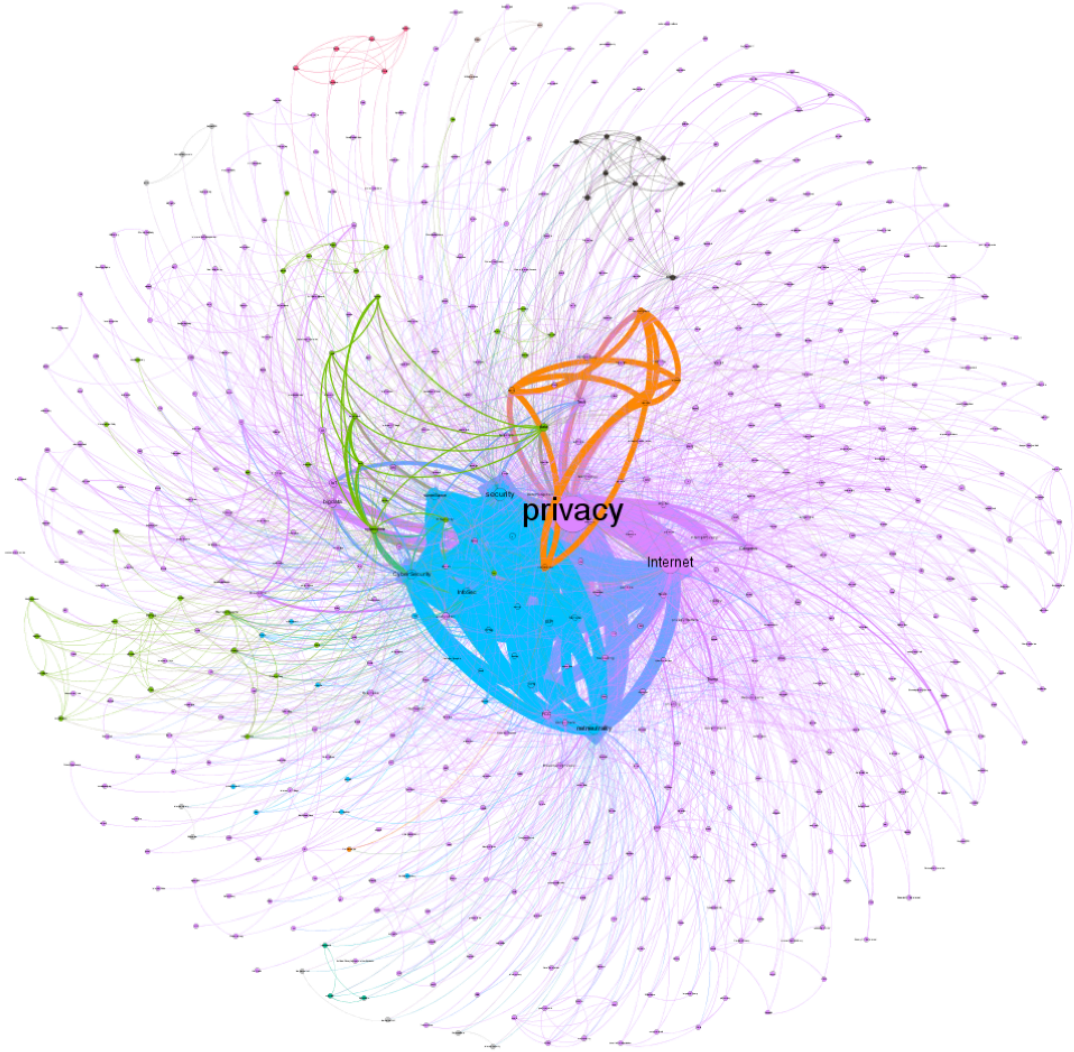
In **light blue** it is highlighted an emerging and interesting **video learning trend**: people sharing their knowledge about a specific topic via the tagging, commenting, and selection of specific frames of YouTube videos, made possible by a service and plugin of Videopin.com. In this case the topics that are "videopinned" are **Windows**, **Alphabet, Google**, **Ibm** and **Watson**.

3.3. Digital artefacts are political artefacts[9]

As we started to see in the previous sections, the discourse around the future of the internet is strictly interrelated with the discourse on its socio-economic impact potential. Interestingly, as particularly evident from the analysis of the "Economy" topics, this discourse is also strongly political: users discuss and express their hopes and concerns about the role played and to be played by public institutions, governments and international organisations when it comes to protect and empower people – or, vice versa – to betray their trust by manipulating and controlling them. At the same time, internet technologies, and social media and big data analytics in particular – are seen as a serious threat to the democratic process. In the next sections, we'll take a closer look to the main issue debated online around internet technologies socio-economic impacts in the near future.

[9] L. Winner (1980): Do Ar3efacts Have Politics? (Quoted by Primavera de Filippi, Next Generation Internet Summit, Brussels, 7 June 2017.

3.3.1. Privacy



Graph 3-10 The relations of the topic “Privacy”

Privacy emerges as a major, cross-cutting, concern. It is tightly connected to major topics and the emotional expressions vary from confident (either because subjects are confident that current negative scenarios will not change, or because they are confident/sure that major interventions and regulations are necessary) to open fear and nervousness about the current situation and its likely future developments. Positive expressions are few and far apart, and, usually, they represent perceptions of business opportunities brought about by the possibility to develop tools for privacy protection.

The topic is tightly connected to the Net Neutrality one, in the sense that a prejudice to the users' online *Privacy* is perceived as a direct attack to the first principle of the Internet, i.e. *Net Neutrality*. The Net Neutrality topic is found in the **blue cluster,** which in turn is strongly connected to the topics ***Data Security***, ***Cybersecurity*** and other personal ***Data Protection*** topics. Other emerging topics are: ***Encryption***, ***Ransomware***, ***Malware***. There is much concern around the fact that Internet Service providers (**ISPs) in the States are now allowed to sell users' personal data** and a **consequent surge of interest in systems like VPNs** for protecting one's privacy. In this regard, it is interesting to note that very wide-spread user data protection systems, such as ***Tor***, are barely visible in the graph. One would have expected more links and conversations about these technologies for anonymous Internet browsing in the Privacy focus. Instead, conversations on this subject are mainly about the Tor browser and the FBI's initiatives against its users or its 24 hours blackout planned for September, 1st 2017.

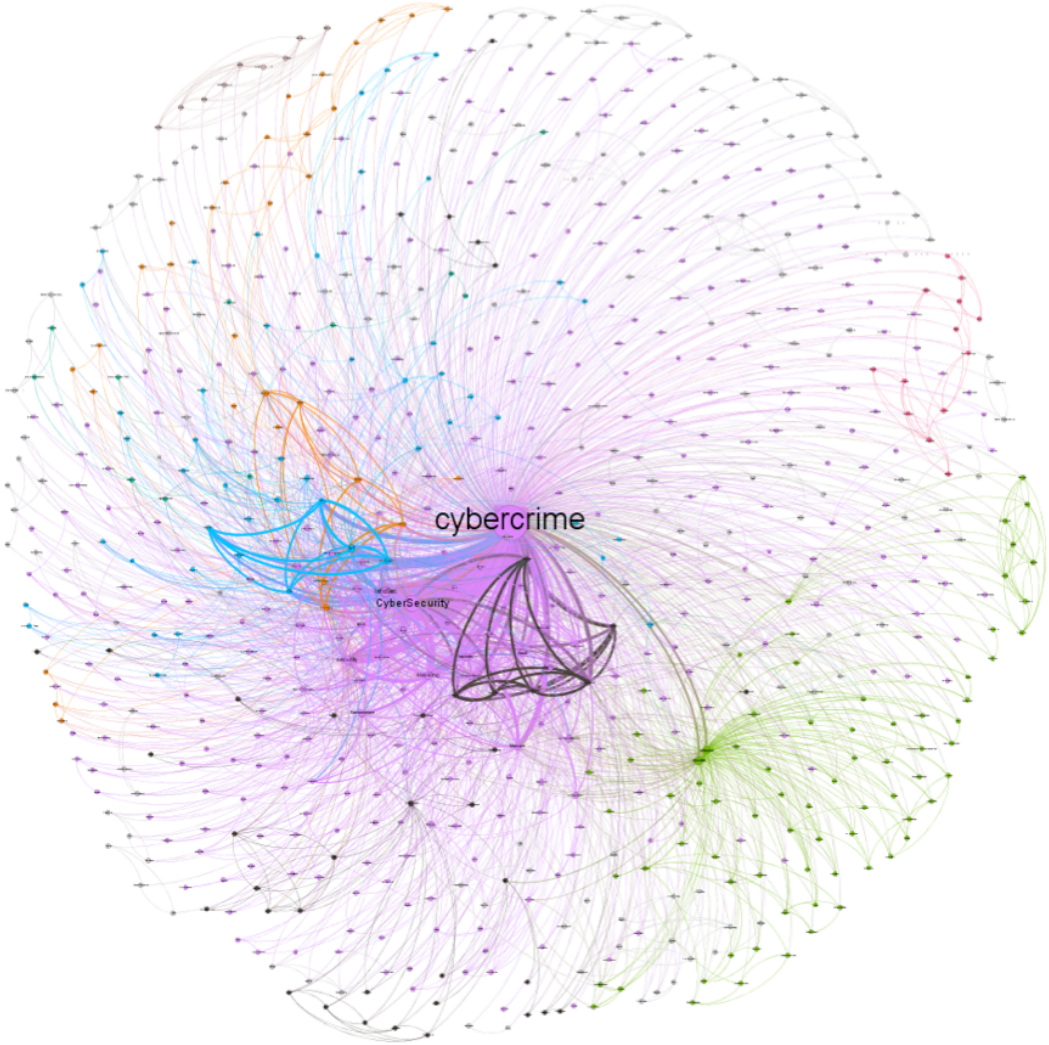
The political and regulatory discourse is very central also in the **purple cluster, showing** strong links with the topic **Congress.** Indeed, the American chamber can strongly compromise the Net Neutrality principle by dramatically changing the **FCC role;** and, on the other hand, has already weakened the privacy regulation approved by the Obama administration. The concern among users is so high that many Reddit and sub-Reddit discussions have been started. Echoes of the past elections and referendums are also often cited as perils for Privacy (for example through the BigData practices connected to elections, for which news and articles shared online often become hubs for active discussions).

In the same cluster are topics like ***Bigdata and IoT***, seen as massive personal data collectors technologies, which might have great impact on users' privacy. We also find many **business companies brands**: indeed, users of brands such as ***Google*** and ***Facebook*** seem concerned about the above-mentioned FCC legislative innovation and the increased possibility for private companies to use their Personal Data for business purposes. See for instance the explicative connection arch: **FCC - consumer - Facebook - information - business**.

In addition, multiple domains are systematically associated to privacy concerns, including health, traffic, domotics, IoT, Apps and Cloud.

The **orange cluster** is mainly focused on the diffusion of the 2018 **Data Protection Regulation**.

3.3.2. Cybercrime/Cybersecurity



Graph 3-11 The relations of the topic “Cybercrime”

Cybercrime – and its opposite, Cybersecurity - is a very rich topic with a wealth of different types of expressions.

The central **purple cluster** clearly shows this variety and, on top of that, the expression of the existence of **a perceived warfare**, composed of **weapons** (viruses, phishing, ransomware, trojans and other, many, malicious software agent types) and **defence mechanisms** (antivirus, scans, news and updates, cryptography and more). In all of this, there is the distinct expression of **subjects which embody the threats**: they can be Russian or generic hackers, deep or dark web participants, scammers, bots and other forms of software agents; the only thing that seems certain in subjects' expressions is the fact that "they" exist, as a recognizable entity which can be blamed, arrested, etc.

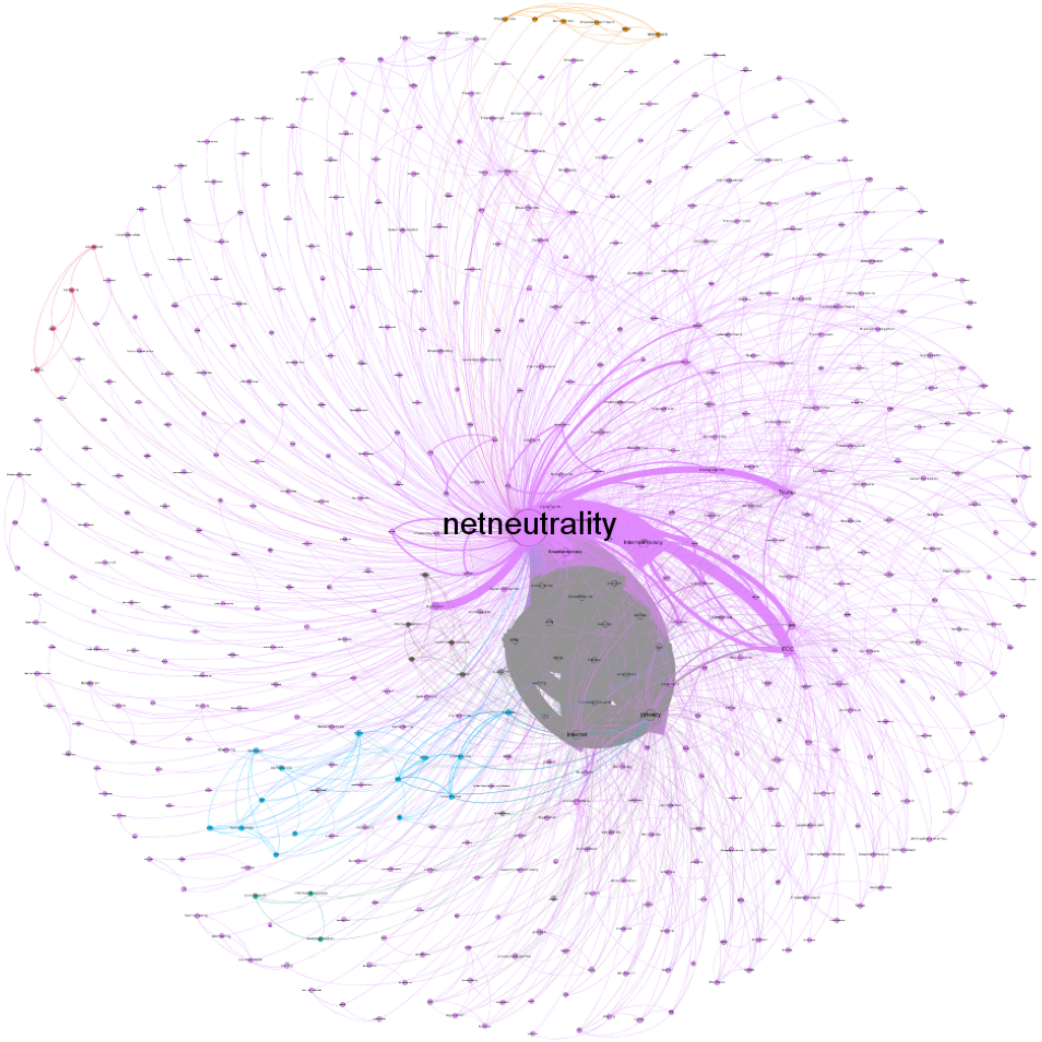
As is shown in the **purple cluster**, the **prevalent emotion that emerges is fear**, and particularly fear of possible ***Attacks***, such as ***Malware***, ***Ransomware***, ***DDoS***, ***Phishing***, etc.

Another fear comes from the ***Fintech*** sector, as the graph highlights its connection with the topics ***Hacker***, ***Cybersec***, but also with devices such as ***iPhone***: the idea is that as access to financial services become simpler and easier, it also becomes more fragile and vulnerable to attacks, and the same apply to other vital services such as the Cloud and other data ***Storage*** systems.

In **light blue** we find strong connections between ***Email, Social Media, Facebook*** and ***Twitter***. With a more in-depth analysis, we found that these topic are related among them and with the security topic since **one of the main security issues is caused today by the fact that people tend to use the same password for different service logins**: e-mail, Social Networks, Bank and Financial services. For different Social Networks the ***Scammer* issue** persists: fake account performing *Social Engineering Techniques* and accessing to users' *Personal Data*.

Global geopolitics are constantly highlighted together with Cybersecurity, for example terrorism, NATO, Daesh, ISIL, and more. Interestingly enough, the word "collaboration" is hardly mentioned in all these conversations which, since multi-stakeholder collaboration has been discussed in many fora around the world as the best approach to tackle cyber-crime, is quite surprising (and slightly worrying).

3.3.3. Net Neutrality



Graph 3-12 The relations of the topic “Net Neutrality”

As described in other sections of the report, **NetNeutrality is a heated topic for discussion, but for a limited number of subjects**. Furthermore, **the typical terms and concepts** which were used to discuss NetNeutrality seem to **have undergone a deep transformation**.

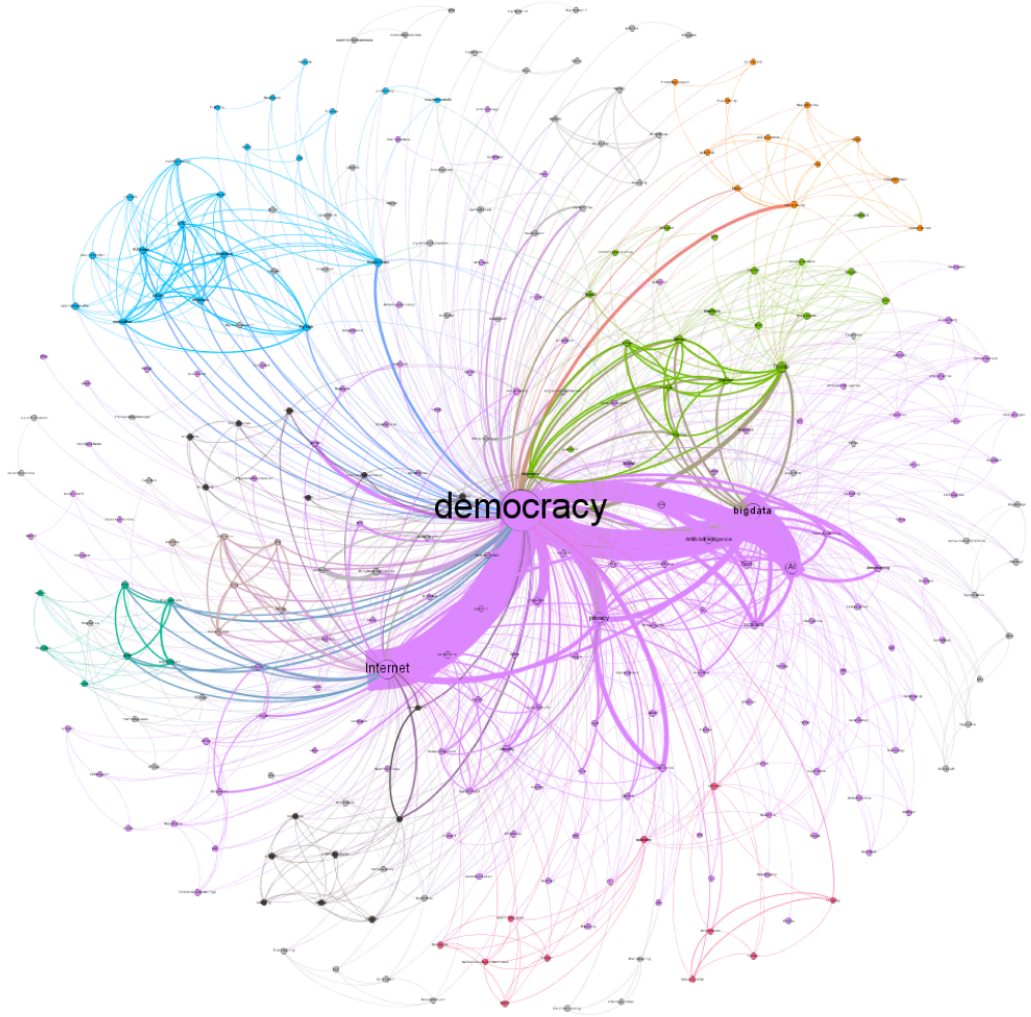
In the analysis, NetNeutrality is nearly completely associated to the entities which, in the subject's opinion, are threating the Net Neutrality principle itself. It might be Trump or operators like Youtube or Netflix, the FCC or the Republicans, and many more. What is certain in the captured expressions is that **NetNeutrality matters**, it is considered a right to be protected and promoted by subjects like the EU and UN, and that it should be the object of resistance and of "No Compromise".

In the main **purple cluster**, the principal concept associations are found. Going beyond the ones previously mentioned: European Union, US and their legislations; the concept of Big Brother and many of its technical manifestations; the emergence of monopolies in media and services; the role of operators (for example telcos); peer-to-peer and decentralization technologies; and Browsers as well as other network access mechanisms, software, devices.

Moreover, in l**ight blue** are the BigData implications in terms of NetNeutrality: domotics, utilities, IoT, are all seen as being potentially connected to NetNeutrality and Privacy concerns.

In **Orange,** at the border of the graph, are the worries expressed by fringe subjects that NetNeutrality issues could be related to the emergence of propaganda schemes, news control and the emergence of Shadow Governments.

3.3.4. Democracy



Graph 3-13 The relations of the topic “Democracy”

Discussions on Democracy are an active playground for the Activist profiles, and revolve around several topics, mostly negatively connotated, such as Censorship, Control, Surveillance, Data, Artificial Intelligence, Algorithms, Science, which are perceived as potential threats to Human Rights, Freedoms and Liberties.

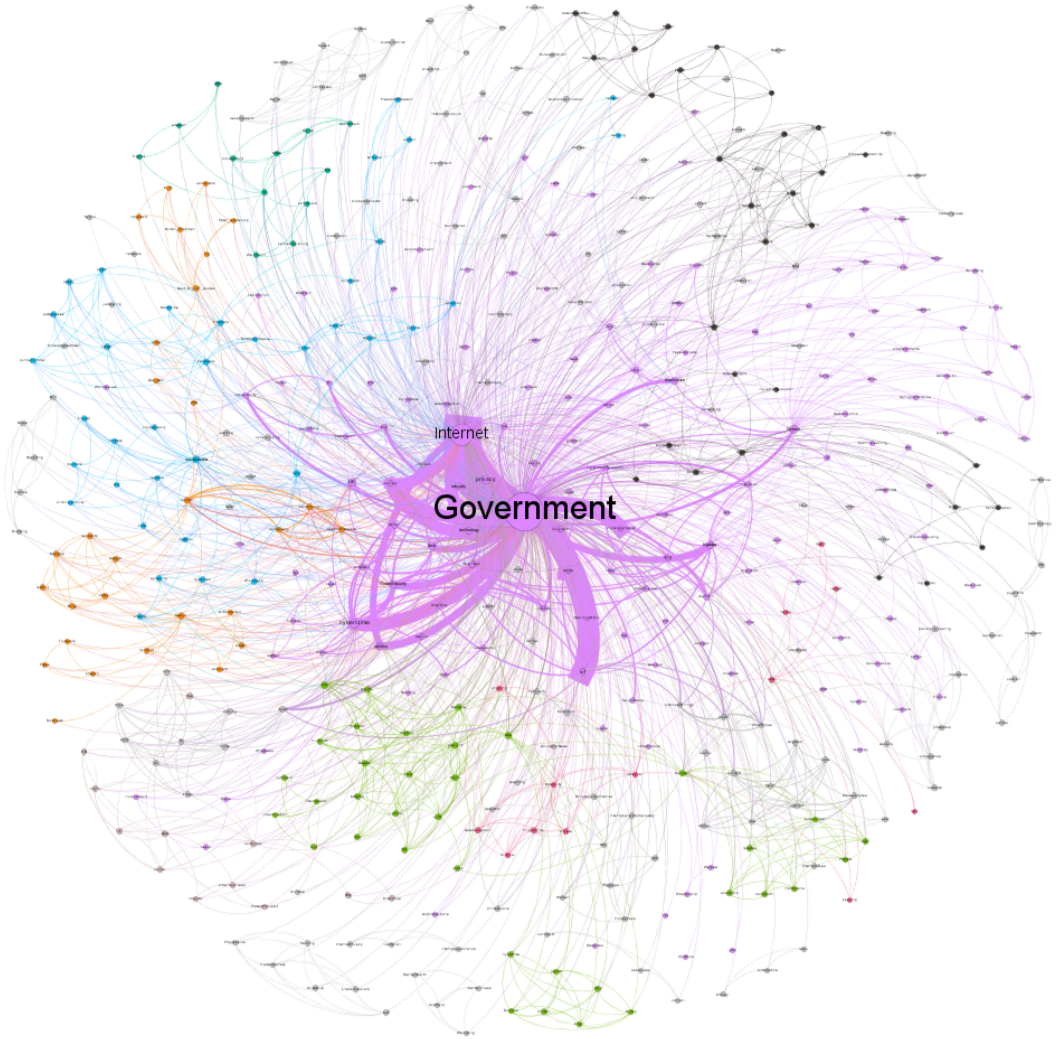
The ***Democracy*** **discussion is highly influenced by the** ***Big Data*** **topic** (purple cluster). It seems that users are wondering if, and to what extent, Big Data analysis systems have actually influenced recent international elections (i.e. US Presidential elections, Brexit, etc.). This is quite clear from the connections between the topics: ***Internet*****,** ***Big Data*****,** ***Privacy*****,** ***Social Media*** **and** ***Election*****, Marketing, Campaigns,** ***Sociology*****,** ***Consumerism***.

Directly connected to this topic is the ***Fake News*** one, in turn well connected to two leading political actors: ***Trump*** and ***Farage***. This suggests that some people are connecting the controversial topic of *Fake News* with some specific political actor or electoral event. From the topic connections, moreover, it seems that people think that the *Fake News* issue should be addressed with some actual, and sometimes extreme, action: for instance, there are strong connections with topics such as ***Education*** and ***Press Ban*****.**

Also worth noticing are the **recurrent conversations connecting** ***Democracy*** **with** ***Big Brother (light grey)*****,** ***Transparency*****,** ***Wikileaks***: showing how for many users, web and social media monitoring systems might easily become a tool for social control and censorship. These fears are magnified by some very active fringes which span from conspirationists to activists: their members, for different reasons and according to different modalities, consider big Internet companies as colluded with Governments, giving access to Big Data and to their monitoring systems in order to activate social control policies. For this communities, the hero for the preservation of a democratic system is Assange and his Wikileaks website, especially when he published the CIA confidential documents on mass control (see for example the following tweets: “Corporate collusion and censorship still exist #bigdata and #bigbrother”; “ Wikileaks: The government is violating your rights and you have no privacy”).

The blue cluster, where the main links are **BlockChain**, **Fintech** and **Democracy**, derives from a **viral Wired article: A Plan to Save Blockchain Democracy From Bitcoin’s Civil War**, reporting about the discussions of the Bitcoin users’ community on the possibility to split the Bitcoin cryptocurrency into two segments, eliminating de facto the peer and non centralized system on which it was based.

3.3.5. Government



Graph 3-14 The relations of the topic “Government”

Analysis of the **Government** topic shows a dichotomy: on one hand "Government" is at the centre of several concerns, not to say fears, while on the other hand it is perceived as a key actor to s**upport innovation and Digital Transformation**.

Legal, regulatory and moral issues related to Privacy, Security, NetNeutrality and Cybercrime are the topics raising most concerns. News items, cases and practical examples are used to express the need of urgent intervention or to aknowledge the inadequacy of current governments.

The **second**, more constructive discourse is more varied. It may deal with the domain of Fintech, in which digital markets, digital payment systems and other similar concepts are expressed to be in need of both regulation and support. Or it may address concepts related to BigData and IoT, which are also expressed, together with AI and blockchains, as the foundations of GovTech and innovation schemes in the near future.

A definite area of discussion (**Orange**) sees Governments connected with rights, freedoms (of speech, for example) and censorship, as well as with private actors such as Google, large media operators and social networking operators whose power should be restricted according to a large number of users.

Multiple **malicious schemes** are expressed as related to Governments. For example, Phishing, which is deemed to be in need of interventions, and also expressed as potentially connected to governments to actuate data collection schemes.

Also worth noticing is the fact that IoT is a “bridge topic” towards **purple and orange clusters** on **Cybercrime** and **Cybersecurity**, all controversial topics, and especially in the UK where there is an ongoing agreement between the UK Government, the entertainment industry and the Web Search industrial giants aimed at the definitive “pirate site” obscuration via advertising revenues cancellation. This is depicted through the links between the topics **Media**, **Google** and **Bing** (see also this link).

The **blue cluster, revolving around the Social Media** and **Facebook topics**, is also very interesting, bringing together apparently conflicting topics such as **Safe Spaces** and **Freedom** on one hand, and **Criminal Activity**, **Infiltrate**, etc. on the other hand. The possible deduction is that Social Networks are seen at the **same time** as **public spaces of maximum freedom of expression, and as media spaces which are often connected to dangers, censorships and other issues which are potentially harmful for democracy**.

4. Conclusions: policy indications and areas for further research and experimentation

Two types of policy indications emerge from the analysis which seem particularly important, the first one concerning inclusion and participation of different audiences to the debate around the FoI; and the second one calling for action in certain areas.

4.1. Indications for inclusion

As we saw in the first section of this report, women and young people are hardly expressing about the FoI, meaning that their voices, ideas and needs risk to have little influence on how the internet will develop in the next future.

Concerning young people, we believe that **specific stimuli should be enacted to facilitate their engagement and raise their awareness about the importance of these topics**. Judging from the results of the analysis, these stimuli could include actions using **design, arts, or gamification processes** and, in any case, serious considerations about the **visual, linguistic and iconic styles** to be used in communication.

Concerning women, our data highlight the extent to which discussions about Artificial Intelligence, BigData, Machine Learning, Robots, Fintech and all the technologies which are likely to fundamentally orientate the development of the Internet in Europe and beyond are dominated by men, with women being relatively active mostly in less technical discussion about education, healthcare and the changing job market. Even in this case, more research into women specific languages and linguistic patterns, as well as preferred channels and topics of expression, would be needed, together with targeted communication and engagement actions.

Another issue which deserves attention and intervention concerns i**nformation and communication bubbles**, whose existence and extent clearly emerge from the extreme fragmentation of online conversations. In this case, it would be necessary to stimulate larger, shared, public discussions, and to make sure results of these discussions are widely disseminated among participants instead of being use by corporations for marketing purposes.

4.2. Indications for action

These indications concern hypotheses for action which strongly emerge from the analysis of the harvested expressions.

Security (Cybersecurity, Cyberattacks, Malware...), **Privacy** and **Surveillance are clearly areas which requires attention from both private and public stakeholders**. The analysis shows not only that citizens expect these **topics as further in need of being addressed, but also as opportunities for the development of new businesses,** where the value proposition would be to protect people's security and rights, putting innovation at the service of people wellbeing. **If matched with the fact that Europe has today the most advanced set of Privacy and Cyber security laws and regulations in the world,** these indications point towards the opportunity to create a serious European competitive advantage in software and hardware production, in data hosting and management services, etc.

Connected to this is the discussion about the **Cloud**. Cloud solutions are seen as a great **opportunity for both Education** (for which competences and skills on cloud systems are highlighted as important all over the conversations) **and Business**. But they are also seen as a potential major **threat to Privacy and Security and as a means of Surveillance**. Based on this, interventions could be designed to **support new businesses and social domains which could explicitly offer the best of Cloud and Security/Privacy**, also referring to distributed solutions (which are systematically mentioned as meaningful and potentially effective solutions).

On a different level, useful indications arrive from the analysis of the discussions around **Algorithms, Artificial Intelligence, Machine Learning** and similar technologies, which are seen as crucial for the development of Europe's technological scenario. On the one hand, they are clearly recognized as **dramatic, disruptive opportunities** (for example in Healthcare), while on the other hand they are explicitly considered as **threats** (such as in Privacy and Surveillance). Re-stablishing trust in democratic institutions would be key to unlock both the social and economic potential of these technologies in Europe. At the same time, and symmetrically, making sure that these technologies are not used to profile and manipulate people, influencing their opinions and actions, is also key to re-establish trust in both institutions (public and private) and in internet technologies.

Another key topic is **NetNeutrality**. It is possible to observe two separate phenomena. As described in detail in other parts of the report, the number and dimensions of **explicit discussion** about NetNeutrality **are minimal**. On the other hand, people are mentioning multiple types of NetNeutrality **related issues**, but without referring directly to the NetNeutrality theme and, most of the times, creating dispersed communications which often only refer to the private entities which are the objects of these discussions, under the form of **complaints to subjects such as Netflix, Youtube and the like**. A dedicated intervention in this area would **bring these discussions together, in public,**

shared ways, so that a larger narrative could be achieved as well as a clearer, more transparent and inclusive opportunity to design shared solutions to what will be a fundamental problem in the near future, i.e. to grant access to high quality broadband connection to everybody.

It would be worth to focus on how to use the central role of **Fintech** across all the topics relating to the future of the Internet. Addressing Fintech means, today, creating impact on most of the other topics in discussion. And the opposite is also true: addressing any major topic bears impact which orbits in or around Fintech. This would call for policies and funding programmes **supporting European approaches of excellence in this field**.

The last indication concerns the role and influence of Europe in building the FoI. From our analysis, it is evident that the debate about next generation internet technologies and their impact is strongly influenced by the American debate. Education is a typical case, but many of the principal topics of discussion have strong tendencies in mimicking and following the tones and styles of discussions which come from the US and – to a lesser extent – from Asia. This is, at least in part, a good thing, since the internet is a global space, thriving to create a shared language between human communities. However, it would be worth considering to stimulate a more “Eurocentric” debate, leveraging on our strengths and priorities to unveil sectors of "Future Design" and Innovation processes which are typically European, not mimicking others.

5. Annexes

5.1. Methodological, technological and ethical approaches to social networks analysis, data extraction and visualizations in REISearch 2017

By including interactive visualizations and social network analysis, this project aims to explore and communicate how European citizens - which massively populate the new public sphere of social networks with their daily interactions - express, discuss and feel about the Future of Internet, to:

- **create a public visual experience** of the data observed and analysed;
- **socialize the data** and make them understandable by a large, non-technical, audience;
- **enable citizens to better understand** the data they produce and their "digital" public sphere;
- **highlight the social and anthropological aspects** of the phenomena we are researching, starting from and including **emergent expressions**;
- **use and explore the potential of big data** in the research process.

Dealing with data – and bringing them into the public sphere – is key. Every day of our lives each of us generate progressively growing amounts of digital data: by shopping, expressing on social networks, exchanging messages, and even by traversing the spaces of the city, using our mobile phones and using our appliances and devices in our homes, offices and schools.

<< This information has started to constitute a large part of our public, private and intimate expressions >>

The publication of the visualizations and the analytic report is part of a **data socialization process**, completed by **a full opendata set** fully accessible by citizens and organizations (see section "Data" and "Licensing and Contributions" for more detail).

Due to the pervasive role of technologies in contemporary societies, **openness** and **transparency** on the methodological, technological and ethical aspects of the processing and analysis of data is here considered not only a formal duty, but a need and a value.

For this reason, this section of the document fully describes:

1. the technologies, methodology and the overall harvesting and analysis process;
2. the technologies and techniques involved;
3. the critical issues;

4. the ethical issues.

5.1.1. Technologies, Process and Methodology

Data is extracted, analysed and visualized by using *Human Ecosystems*, a set of techniques, technologies and methodologies elaborated by HER - Human Ecosystems Relazioni[10].

Step One: The Harvesting Process

The first step is a harvesting process, in which major social networks are monitored in order to detect public content generated in Europe about the Future of Internet.

Sources:

public contents from Twitter, Instagram, 500 selected Facebook public groups and pages of thematic interest

Capturing content from the various sources requires different techniques.

[10] HER - Cultural Acceleration, through Open Big Data. For more information: https://www.he-r.it/welcome-to-her/

For example, services like **Twitter and Instagram provide APIs** (Application Programming

Interfaces) which allow searching for certain keywords, hashtags, geographic locations and timeframes, and, thus, to obtain the public content which was generated by users about certain topics and in relevant locations. There are limits for the usage of such APIs (for example on the number of contents which can be harvested, on the geographic area which can be searched, on the amount of time in the past for which it is possible to perform searches, and on the overall usage of the APIs themselves). Nonetheless, by combining the available data access points and modalities, it is possible to explore thoroughly the public content generated regarding specific topics and in specific locations.

Other services, such as **Facebook**, are much more restrictive. A series of APIs and frameworks (for example the Open Graph and a few other ones) are available, but the limits for their usage are much more stringent. Content is obtained by performing an initial search for those pages and groups that are relevant for the collection process, directly connecting to them (for example by using the "Join Group" function available on the social network) and, then, using the APIs, through which it is effectively possible to monitor the content which appears on them.

On top of this, all of **the services impose limits about how the harvested content can be used**. For example, it is not possible to store it directly in databases; when using it is necessary to provide the indication of the links from which it originates; it is necessary to provide attribution and

declaration that the use is noncommercial, and similar ones. Through HER's solutions it is possible to have all of these requirements satisfied automatically (they are specified in the various *Terms of Service agreements documents*, for users and developers, available on the respective social networks' websites).

Step 2: Processing the Data to generate Knowledge

The **collected content is, first, anonymized and aggregated to form clusters which are suitable for analysis**, around different logics (for example by forming groups of contents by counting the mentions of a certain keyword).

<< The content is stored in databases only in this form (anonymized and aggregated) >>

This is a very delicate stage, as it implies the verification of multiple types of conditions which not only ensure proper anonymization, but also the fact that, given an anonymized content, it is impossible (or really, really difficult) to climb back up to the original, identifiable, one. (For example: even though it is made anonymous, a geo-referenced data which is alone and isolated on a territory may make it too easy to understand to whom it refers to; this is why we remove these singularities and other similar cases from our databases).

In general:

HER's systems are fully compliant with current EU regulations on privacy, personal data collection and management, and anonymization of personal data. On top of that, a dedicated team at HER monitors changes in laws and regulations to make sure that this fact remains persistent.

When this critical phase is complete, the data is processed to generate knowledge.

There are a number of processing and analysis techniques used, such as Natural Language Analysis, Emotional Analysis, Network Analysis, Geo-Referencing. They will be described in more detail in the next section.

In this overview, it is important to highlight how these techniques are able to **transform the unstructured data collected from social network** (messages, images, comments, conversations…) and process it in order to transform it **into structured data**, forming the knowledge base of the research.

At this stage five typical types of knowledge are produced:

- **topics**, as content is scanned for what it is talking about, and for what topics are discussed together in the same contexts;
- **emotions**, as content is scanned to gain understandings about what emotions (such as happiness, surprise, fear, anxiety, disgust, trust…) they are expressing;
- **times**, using both the content's meta-data and the phrases it actually contains to understand what time it refers to (for example mentions of events or recurrences);
- **places**, where the content's meta-data (such as geographical coordinates) or sentences describe geographical locations;
- **networks**, in which the focus is to understand which people, organizations and other entities these contents put together, describing the models of relational networks and graphs (for example: do individuals talk among themselves or with organizations? or: does information come from news items or from peer-to-peer discussions? or: how extended are networks discussing a certain topic? and similar ones which allow to determine the models of communication).

All the information elements are semantically linked with each other and, thus, can be combined to infer more complex knowledge (for example, by combining knowledge about topics, places and times, we could be able to infer what the people discussed at a certain event; or by combining topics, emotions and networks we could understand what kind of communities express which emotions about certain topics).

The content of the knowledge base is also used as a feedback process, to fine tune the data harvesting process, using a Machine Learning mechanism: here all the accumulated knowledge is used to evaluate new information to generate new knowledge about how to modify the data capture process, in terms of other words/topics to listen to, other pages, groups and communities (for example on Facebook) to include in the capturing process and other insights of similar nature.

The acquired knowledge is used in the following cycles, obtaining a system that learns and adapts to the evolving scenario (for example by understanding that at a certain time it may be interesting to include some other elements in the harvesting process, as they are particularly active and relevant).

The knowledge base is, then, used to perform some more standard analysis, such as qualitative, quantitative and community/ network analysis, to gain better understandings about the scenario that all of this information describes, such as:

- the **timelines** according to which the topics, emotions, places and communities evolve;
- the **topics**, according to which we are able to gain better understanding of how much certain topics are

discussed, with which emotions, by which communities and in which places;
- the **communities**, with which we are able to understand how diverse or coherent different communities are, what they focus on, how they converge or diverge, what are their main concerns or desires;
- the **flows**, using which we are able to model how information, opinion, influence spreads;
- the **impacts**, with which it is possible to gain understandings about the results of certain actions, such as how a communication campaign or even a single social networking message is able to influence people's behaviour;
- the **correlations**, helping in comprehending possible cause/effect relationships;
- the **transformations**, in which it is possible to take the dimension of time into account, to study how all of the above evolve in time.

At this stage an evolving knowledge base is created.

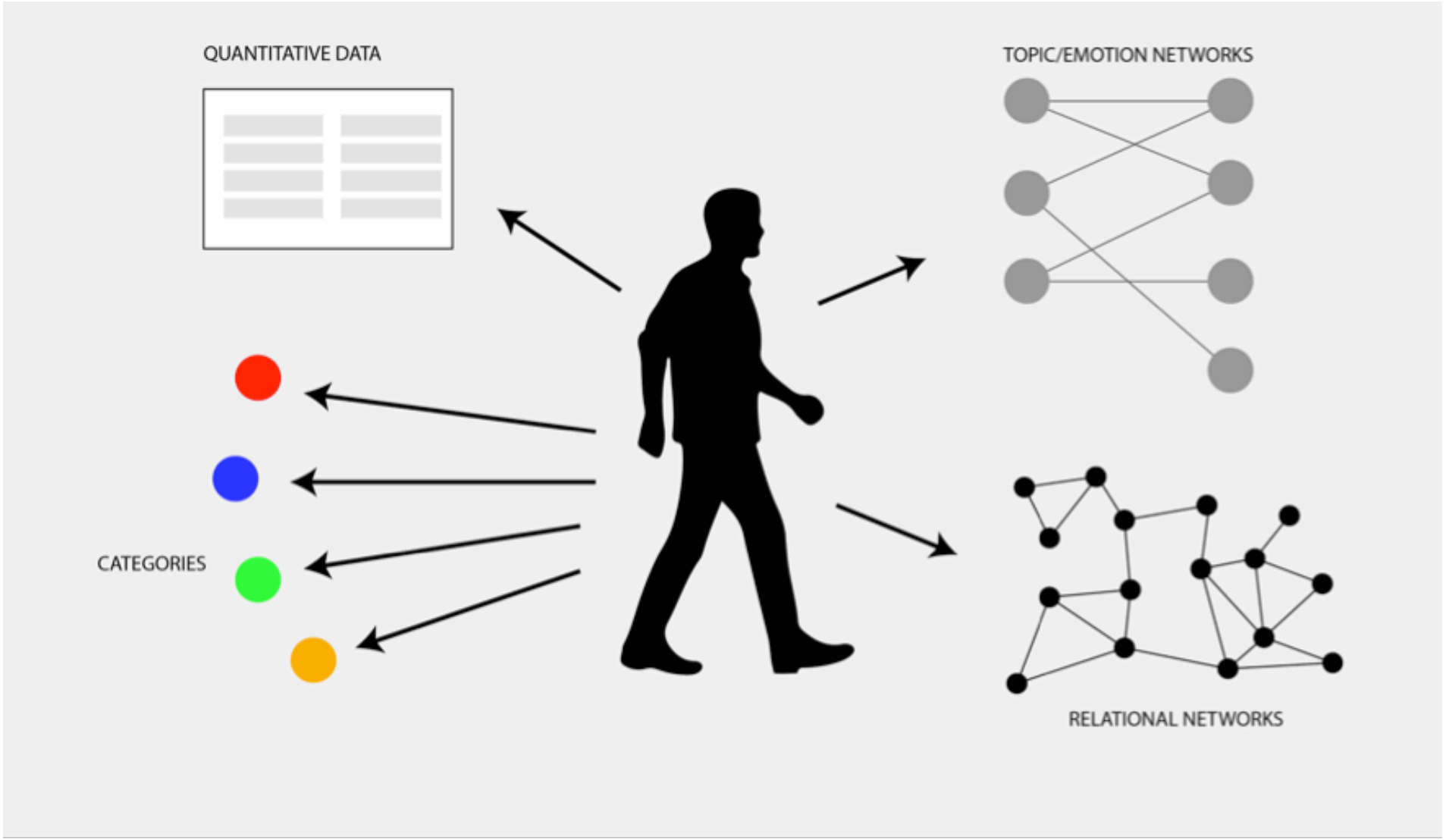
Step 3: Socializing the Data

The final step is the data **socialization process**. In the context of this research, **visualizations and Open Data are the tools** designed to open and enable this process: our **website** is the access point for both.

The interactive visualizations can be browsed, observed and used as an actual resource through:

- the embeddable, a small piece of code which enable to publish the visualizations on any website, just by copying and pasting it. Made available for the media partner during the campaign, embeddables can now be used by anyone to recreate interactive experience
- the opendata set, made available for each visualization and listed in the following annexes.

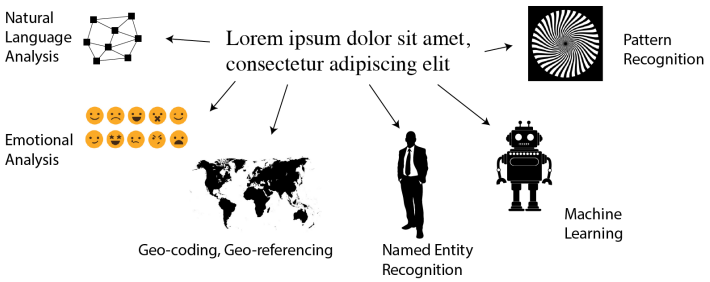
The **analytical report**, too, comes with **a full opendata set**, becoming a resource to be used by civil society, university, researchers, students, designers, artists, journalists and anyone interested in it for their own purposes.



Graph 5-1 Types of Data which can be captured and processed, transforming sensitive personal data into aggregated, anonymized, data)

5.1.2. Techniques and Technologies

This section is intended to provide a short description of the main techniques and technologies used for processing the harvested data.



Graph 5-2 Techniques and Technologies used

Some links are given to obtain further information about each technique for anyone interested to learn more on it.

Natural Language Analysis

The objective of **Natural Language Analysis** (or Natural Language Processing, **NLP**) transforms unstructured data such as text into structured data. It can be performed in multiple ways, with different objectives, such as understanding the topics which a certain text deals with, creating automatic summaries, machine translation and more.

The technologies used in the analysis (see previous sections) employ advanced **NLP, which is performed in 29 languages**, and is used in the following ways:

- **Discourse Analysis**, which deals with understanding the structure of text and its components; for example using the way a certain sentence is written to understand if it is a question, an exclamation, a sentence providing information of some sort, an answer to a certain question, etc;
- **Semantic Analysis**, which deals with starting from text to understand its meaning, in terms of whether it assesses a certain topic and in what way, if it has a certain style for expression or if it uses a certain language;
- **Topic Discovery**, in which large numbers of sentences are observed to discover if recurring patterns may identify new topics to listen to which are relevant for the ones currently being observed; new topics come under the form of words, word patterns, sentence patterns and more;
- **Named Entity Recognition**, which uses streams of texts and their structure to identify proper names for people, places, events and more;
- **Relationship Extraction**, which uses text to identify the relationships between Named Entities (e.g.: who is married to whom; who is the employer of whom; etc.);
- **Sentiment/Emotional Analysis**, in which the words and the patterns in which words are composed are used to gain better understandings about what

Sentiment the sentence is expressing (positive, negative, neutral), or, if enough information is available, what emotion it is expressing (such as joy, fear, anxiety, surprise, trust, satisfaction, etc.);

- **Information Retrieval and Information Extraction**, which, given the procedures listed above, deals with the possibility to store and extract the types of information which can be extracted from text.

More information on NLP can be found here:

https://en.wikipedia.org/wiki/Natural_language_processing

Emotional Analysis

As described in NLP, **Emotional Analysis deals with the possibility to automatically recognize emotions in text**, by recognizing how text uses word, phrase or sentence patterns.

The technologies used in the research allow recognizing 36 main emotions. This occurs when enough evidence is present in the texts, confirming the nature of the expression being analyzed. Classification of emotions is performed using the **Circumplex Model of Affect**.

In general, the Circumplex Model of Affect classifies emotions according to two main parameters: **Energy/Arousal** and **Comfort/Discomfort** (and, optionally, according to a series of further ones). In computationally analyzing the text, words and their combinations are accounted for according to their Energy and Comfort/Discomfort contribution to the sentence and, thus, the total can be used to infer whether a certain sentence is expressing a certain emotion. The levels of Energy and Comfort are determined through large semantic databases which Human Ecosystems, the technology used in the research, has built across the years, in 29 languages, and has been built both manually, in collaboration with people and researchers from all over the world, and automatically, using machine learning to understand when certain systematic recurrencies confirm the fact that a certain expression denotes a specified level of Energy and Comfort.

More information about the Circumplex Model of Affect can be found here:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2367156/

Network Analysis / Social Network Analysis

Network analysis studies **graphs**, **networks of relations** between discrete objects, or nodes.

In the technologies used by the research **Network Analysis is used to study the composition of the networks represented by the people, organizations, companies and not-human agents** (eg.: bots) whose expressions are captured through their public online expressions and, given these and their transformations, the flows of communication, information, knowledge take place in and through them, effectively describing how information, opinion, emotion, knowledge and influence spread across communities and cultures.

Standard network models are used in the research (such as networks described through graphs made up through nodes and links, and stored in relational databases or, more often, in network databases) to work on the relational graphs, while to gain better understandings of their functioning, a custom version of **Latour's ANT** (Actor Network Theory) is implemented to describe the behaviours of networks and of their participants, and to identify roles within them, such as influencers, experts, hubs, bridges among different communities, and the models according to which certain forms of interactions lead to specific results (such as the propagation of a certain information element into a community, or its ability to cause a certain type of reaction).

In General, statistical models and graph-theory models are used to analyse the networks, such as calculation of degrees, network diameters, graph densities, authorities[11], modularities[12], page rank[13], connectedness[14] and other similar ones. The results of these computations and indices are then fed to machine learning algorithms, both instantaneously and as networks evolve in time, to discover recurrencies and patterns, which may highlight the formation of interesting network/community configurations.

For more information about Network Theory:

https://en.wikipedia.org/wiki/Network_theory

https://en.wikipedia.org/wiki/Network_theory

For more information about Social Network Analysis:

https://en.wikipedia.org/wiki/Social_network_analysis

Geo-Referencing

This technique is the process of attributing a geographical context to a certain content.

The **geo-context** can be of multiple types:

- the **location** in which a photo has been shot;
- the **area** for which a certain content is relevant (for example a city or a state);
- the **path** along which a certain information is relevant (for example the path that takes from a certain location to another).

In the technologies used in the research geo-referencing is performed in two ways:

[11] Jon M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, in Journal of the ACM 46 (5): 604–632 (1999)

[12] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000

[13] Sergey Brin, Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, in Proceedings of the seventh International Conference on the World Wide Web (WWW1998):107-117

[14] Robert Tarjan, Depth-First Search and Linear Graph Algorithms, in SIAM Journal on Computing 1 (2): 146–160 (1972)

- **using the meta-data** included with contents, for example the geographical coordinates which social networks users can associate to their posts;
- **using the results of NLP**. In this case the **Named Entities** identified in text may be of geographical relevance (for example the name of a church, or a landmark, or the name of a restaurant); if the sentence includes enough evidence of the spatial character of the expression (for example the sentence may state that "I am going to…"), sufficient information may be present to identify the geographical context for the content. For this research, this modality has been used to identify whether the posts mentioned European nations, regions, cities, or universities, research centers, institutions, organizations. For this, a database of all European Nations and administrative regions, the largest 500 European cities by population, and a list of European Universities[15].

On top of that, a **GIS (Geographical Information System)** is used to establish the hierarchical characteristics of space, according to which certain coordinates are contained in certain blocks, which are contained in certain neighbourhoods, which are contained in certain zones of the city, which are contained in the city, and so on.

15 source: https://en.wikipedia.org/wiki/Lists_of_universities_and_colleges_by_country

Machine Learning

Machine learning (ML) studies the possibility to design algorithms which are able to **recognize recurring pattern**s (in this case: patterns in texts and other parameters, such as quantities, time series etc.) and to use the fact that certain patterns have been recognized to learn, producing systems which automatically adapt themselves to changing scenarios.

In the technologies used in the research, Machine Learning is used in multiple ways:

- for the NLP techniques;
- in Topic Discovery;
- in Emotional Analysis;
- and to fine tune the data harvesting processes by adjusting what keywords, phrases and forms of sentences are monitored on social networks.

Here we proceed in analyzing each of these cases.

ML for NLP

As highlighted in previous sections, ML is used for NLP to augment the semantic databases used for classification of the captured texts.

Semantic databases (such as Wordnet, which is used in this research together with its translations and extensions, including ones which have been produced over time in 29 languages, using the techniques described in this section) are composed by by synsets, which are not single words,

but, rather, sets of words, their relations, short explanations and usage cases.

NLP techniques use these synsets in different ways to try to transform unstructured data (text) into structured data by mapping and matching synsets to sentences and, thus, understanding parts, subjects, objects, verbs, topics, proper names of places, people, events, and more.

In fact there are multiple forms of NLP, including discourse analysis, named entity recognition, topic discovery and more.

Some of these techniques will be analysed in following sections.

Here we will focus on how ML is used in the systems used in this analysis to enrich the vocabulary of synsets, to allow for constantly better semantic analysis.

Text is, first, cleaned and prepared (for example eliminating stop-words and other problematic symbols).

Then it is fed to a discourse analysis process whose output is the estimate of the possible syntactic architectures of the sentences (for example, the same sentence could be interpreted in more than one way, for example due to ambiguous punctuation).

Then both the text and the result of discourse analysis are fed to subsequent processing stages, falling principally in the Latent Vector Semantic Analysis[16] family of techniques, whose output is a classification of sentences and of their parts into lists of topics and categories.

These are used to statistically evaluate the meaning of texts and to extract topics and other classifications from them.

In all of these phases, ML is used by taking the various outputs of the different stages to discover recurring patterns which may indicate the existence of candidates (words, phrases...) to be included in existing synsets or to create new ones.

This type of task is performed by using various forms of neural networks whose target function activates when recurrent patterns are found in the outputs. For example, in discourse analysis, if a certain pattern of words or sentence parts is found to systematically (ie: a high number of times) to produce the same output, it is flagged to be considered as a good candidate.

These lists of candidates are processed both manually and automatically.

Manually: samples are manually evaluated in order to provide positive and negative examples, which are used by the neural networks to change weights in the net's nodes.

---

[16] Strapparava, C. (2010) Semantic Similarity from Corpora - Latent Semantic Analysis, Accessed May 10 2017: http://clic.cimec.unitn.it/marco/teaching/compling/materials/LSA-2010.pdf

Automatically: particularly high thresholds are set to trigger automatic transitions from "candidate" to "official" (for example, if a threshold of 50 was used to flag possible candidates, a threshold of around 500 would be used to transition to official synset).

In this way, the semantic databases are maintained both manually and automatically in very powerful ways, both through extension (positive examples) and contraction (negative examples). Language information (consisting of flags which are found in the databases, as multi-language content is very frequent on social networks, both formally, to provide users with multi-language content, and informally, where words from different languages are mixed) is added to enable language detection and classification, and similar ML processes are applied also to those (for example: if a certain synset is systematically found in french sentences, after a certain threshold it is flagged as "french candidate").

ML for Topic Discovery

Similar techniques as described in the previous section are used on topic classification outputs to discover new topics.

For example, topic lists which have been detected in a certain paragraph can be fed to neural networks configured to react to recurring patterns across the sets. For example, a combination of N different topics could be detected to appear systematically together with similar probabilities (because, we have to remember, that all of NLP's determinations are probabilistic, so a certain sentence would be described as being XY% relevant for Topic A). If this happens and the number of times is sufficiently high, the set is flagged as candidate as "new topic".

This technique is applied also in mixed terms (for example among sets of topics and synsets, not only to topic-topic sets), to detect potential new topics which are not only the result of remixes of already existing topics.

The positive/negative example technique described in the previous section is used also in this case to extend/contract the knowledge of the system.

ML for Emotional Analysis

Similar technique is used to extend the capacities to classify emotional expressions using the circumplex model of affect.

When using Latent Semantic Analysis, synsets are characterised as vectors in multi-dimensional spaces whose dimensions are topics and other logics: for example, a certain synset which, if present in a text, highly contributes to Topic A would have a high coordinate along the Topic A axis. To perform emotional analysis using the Circumplex Model of Affect, axis for arousal and pleasure are present in the multi-dimensional space, so that each synset can be characterised according to its contributions (singularly and in relations) in terms of these two dimensions. This allows for composing the contributions of

vectors to understand the overall characterisation in terms of the circumplex[17].

This means that in this model, certain areas of the arousal/pleasure plane will be associated to different emotional characteristics.

Which, in turn, enables to design neural networks which recognise recurring patterns on this plane. Patterns in particular areas can be, thus, recognised, to make weighted hypotheses about particular synsets participating to specific emotional expressions.

These synsets are flagged as possible candidates. Then the positive/negative example scheme and the high-threshold driven, automatic, mechanisms are used to augment knowledge about emotional classification in the system.

Other Techniques

A few additional techniques have been used in this research:

- to **detect demographics**;
- to **detect the type of account** (individual, organization, bot).

[17] Bellegarda, J.R. (2013) Data-Driven Analysis of Emotion in Text Using Latent Affective Folding and Embedding. Computational Intelligence, Volume 29, Number 3, 2013.

To detect demographics we have used research coming primarily from health-related practices, to be able to use approaches and techniques which have been tested and evaluated in strict, formal environments, and to be able to build on medicine's ethical approaches, which are among the most stringent in regards to preserving people's rights.

For this to infer gender and age group from the harvested posts we have followed the indications emerging from the following study:

- **Cesare, N., et al** (2017) "*Detection of User Demographics on Social Media: A Review of Methods and Recommendations for Best Practices"*, arXiv:1702.01807 [cs.SI]

The study conducts a thorough review to evaluate the different modes in which researchers have documented to have extracted demographic data from social networking posts of various kind.

In this research we have selected and combined modes indicated in the paper as "scalable", in the extraction of "gender" and "age group":

- **M. Vicente, F. Batista, and J. P. Carvalho**, "*Twitter gender classification using user unstructured information*" in Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on, 2015, pp. 1–7

- **P. A. Longley, M. Adnan, and G. Lansley**, "*The geotemporal demographics of Twitter usage*" Environ. Plan. A, vol. 47, no. 2, pp. 465–484, 2015
- **J. Chang, I. Rosenn, L. Backstrom, and C. Marlow**, "*ePluribus: Ethnicity on Social Networks.*" ICWSM, vol. 10, pp. 18–25, 2010

In our case, from a sample of 2000 random elements, the application of these algorithms has proven to be **82% accurate**.

To detect the type of account (for example, whether it is an individual, or an organization or a bot), results from these publications have been used:

- **Varol, O. et al** (2017) "*Online Human-Bot Interactions: Detection, Estimation, and Characterizatio*n" arXiv:1703.03107 [cs.SI]
- **Davis, C. A. et al** (2016) "*BotOrNot: A System to Evaluate Social Bots*" arXiv:1602.00975 [cs.SI]

In this sense, the frequencies and characteristics of the posts and of the words and tags and shares which their corresponding social network users have expressed, are used to determine a 0-100 score in which 0 means "human", and 100 means "bot". In the 40-80 range value, a NLP technique has been used to detect use of collective, impersonal or other "non individual" means of expression and self-representation, to attempt at detecting "organizations", and differentiating them from both humans and bots.

From a sample of 2000 random accounts tested from the ones which have been collected, the results have proven to be **74% accurate**.

5.1.3. Critical Issues

A series of critical issues have been identified during the research process. The following are the most relevant ones.

Privacy Laws and Regulations, and Terms of Service Documents

Most online operators provide their services under condition of accepting their **Terms of Service** (ToS) documents. These are legal documents which Internet Users acknowledge and approve when subscribing to these services.

Social networks have very complex, strict ToS documents, which are intended both to preserve people's rights while they use the online platforms, and to ensure that the business interests of the providers are protected.

Currently different ToS are provided for both regular users and for developers, specifying different levels of access, usability, availability of the data and functions made available by the different service providers.

<< Limits described in the ToS documents include the ways in which information from these platforms can be extracted and used. >>

At the same time, a complex set of Intellectual Property, Data Access and Usability and, even more important, Privacy and Data Protection laws and regulations exist at multiple levels, for example National and European.

It has been published in May 2016, will be active starting May 2018, and the project is already fully compliant with it.

Establishing how the policies of Social Network providers match and are compliant to the EU policies is very difficult matter. Even though a certain amount of transparency is mandatory, the presence of proprietary technologies and the frequency at which the technological systems of these providers mutate (for maintenance, ordinary source code updates, interface changes, etc), i**t is virtually impossible to thoroughly describe how (and if) these systems comply with what the EU** has described as mandatory to preserve people's rights.

At the same time, **putting the ToS documents and the EU regulations side by side lets a plethora of grey areas emerge**, and opportunities for "interpretations" which cause this type of activity to be far from certain both in its actions and effects.

In the context of this research we consider of great importance both to preserve people's rights, according to EU laws and regulations, and to preserve the needs of businesses and organizations.

In this sense we have decided to systematically apply EU laws and regulations, first, and, then, to apply what is dictated by the ToS agreements coming from the various service providers.

As described in the previous paragraphs, there is a level of opacity and of complexity which causes the impossibility to fully ensure respect both EU laws and regulations and ToS documents, at the same time. While EU laws and regulations are "open" and "transparent", ToS documents are based on systems which are opaque by nature and design (proprietary, closed source). Thus it is not really possible to understand how their "clear" and "transparent" text corresponds to actual practices.

Considering the object and the nature of this research, inquiries, discussions, code analysis, and open sourcing of data and source code are welcome to be then analyzed in open, public, transparent processes, with all stakeholders and interested parties.

Quality of Automatic Interpretation

This issue deals with the **quality of the interpretation** of the content as processed by the automatic algorithms.

This means to try to ensure that if algorithms detect that a certain content deals with topic X, the content effectively deals with topic X. This is a very complex thing to do. While performing tasks such as NLP, algorithms are de-facto collecting bits of evidence across texts, such that at a certain point enough evidence can induce us to believe that a certain content is effectively talking about X. But these are not final determinations, they are probabilistic: for any combination of such evidence, we will be always XY% sure

about this fact, and XY% will never fully be 100%, there will always be a doubt.

The technologies used in this research confront with this type of issue by **establishing very high thresholds**. Currently the systems used accept a certain interpretation only if there is evidence to prove it which accounts for **more than 95%** of probability.

Irony

This issue is a peculiar version of the previous one.

Social media (and Internet in general) is a context which is characterized by high degrees of irony. This means that the situation in which someone is expressing something and really meaning its opposite will happen very often. In computational terms, this means that the situation in which an algorithm will efficiently identify topic X or emotion Y in a message and the user generating it meant the exact opposite (or similar situations), will happen very often. This is currently one of the most pressing issues in Natural Language Analysis: Irony.

There are a number of techniques which are currently used to mitigate these issues. All of them take into account the context in which each message is generated.

By studying the context in which a certain user communicates (their beliefs, opinions…) **we will have better tools to interpret an ironic conten**t. This is what is performed in this research.

While user accounts are anonymized, they are processed using the Topic analysis techniques which have been analyzed in the previous sections. If for a certain topic at least **75%** of one users' expressions is polarized in a certain way, a further expression which is polarized very differently will not be accepted immediately, but placed in a limbo, "on hold", until enough further evidence will be able to prove that the user has changed opinion. This means that these 'limbo' expressions would be excluded from further aggregations and calculations, because there is not enough evidence which allows to interpret them with sufficient quality and certainty.

Lack of Intentionality

With this issue we refer to the possibility that online expressions do not always reflect what online users chose to express, with intention.

This is an issue with multiple faces. For example, by understanding a certain message we could be able to collect enough evidence about a person's behaviour, or opinion. This fact is only partially related to the same person's belief system, or values, or desires. The person might have been angry, or in a hurry, or even forced to express in a certain way, for respectability, reputation, work, shyness, or multiple other reasons. Or, on the other hand, people may not consciously realize that they are debating issues in the public sphere. Or they may even not realize that they can establish such debates, and not talk at all about such issues, even if they care about them.

In general, little can be determined about the intentionality of the expressions, due to their emergent, informal character.

5.1.4. Ethics

Many of the issues identified in the previous section can be merged together with other, more general, ones in defining the need for the composition of a comprehensive ethical code.

These are the principal elements the Code of Ethics and Conduct adopted in the present research:

- full respect and compliance for recognized laws and regulations, at regional, national, European and international levels;
- explicit and avoid conflicts of interest of any form, especially for whatever concerns the code of ethics and conduct;
- provide clear and accurate communication;
- operate with transparency and integrity;
- protect people's data and rights, by respecting current laws and regulations and by providing full access to data, information and knowledge.

5.2. Open Data

This annex includes the indications about where to find the different types of Open Data made available by the research.

Open Data are provided in 2 different formats:

- **CSV**, as in Comma Separated Values, which is a textual format for data, encoded in fully internationalized UTF-8 format, and which is accessible through commons spreadsheet software, such as the ones found in Open Source suites like Open Office, Libre Office and more;
- **JSON**, which stands for JavaScript Object Notation, which is a different textual format which employs JavaScript's Object Notation to represent data as objects and arrays of object; it is common to use JSON formats to represent hierarchical data, and to use it in software programs written in most programming languages (such as JavaScript or PHP, for example).

Below is the list of open datasets provided with the research:

5.2.1. General Datasets

Title	Description	Format	URL
Comparison	Provides the data about how many people were discussing and how many messages were exchanged about Future of Internet, Brexit and Soccer, during the time of the research and using the same sources.	JSON	comparison.json
Demographics	The age groups and gender of the subjects analyzed in the research	CSV	demographics.csv
Emotion Types	The data dictionary holding the 34 main emotions classified by the technologies employed in the research, with each also indicating the corresponding central level for Comfort and Energy, according to the Circumplex Model of Affect	CSV	emotion_types.csv
Emotions	The principal emotions for the top topics discovered in the research. Each line provides the topic, the weight of the topic, and the average Comfort and Energy expressed for that topic	CSV	emotions.csv

Emotions for Subjects	Indicates how many subjects expressed certain emotions. Each object in the array includes the “c” field which contains the number of users which expressed the emotional expression indicated by the “comfort” and “energy” field.	JSON	emotions-subjects.json
Flows	It describes the flows of communication among the types of subjects (see “Types of Subjects” dataset). The “nodes” array includes the list of the types of subjects. The “links” array holds objects which specify ordinal number for source and data node for the link (if “2” is in the “source” field it means that the source of the link is the second item in the “nodes” array) and the “value”, indicating the number of communications occurring between source and target subject types.	JSON	origins.json
Languages	The languages detected in the research. Each row indicates the name of the language using both its 2 letter ISO code and its full name, and the number of times in which it has been detected in the research. The “short” version combines languages with less occurrences into a single “Others” item	CSV	languages-short.csv languages.csv

Profiles/Segments	As described in the previous sections, Machine Learning has been used on the top topics to segment 3 main profiles (Optimist, Activist, Exploiter): these are the typical emotional expression of each profile. Each row includes the indication of the profile, a topic and the level of typical Comfort and Energy.	CSV	profiles-data.csv
Profiles/Segments Descriptions	They are the descriptions and meta-data for each profile. Each object includes the name of the profile, its average age, its quantity in the research, and its description in HTML format.	JSON	profiles-descr.json
Topics	The top topics discovered in the research. Each row has the name of the topic and its weight, describing how many times it appears.	CSV	topics.csv
Types of Subjects	The types of subjects detected in the research. Each row includes the indication of the type and how many were detected.	CSV	types-of-subjects.csv
Topics Network Overview	These are the nodes and links of the general network of topic relations. The nodes files indicates the nodes of the network: the list of topics. Each row includes an ID for the topic, its name label, its weight (the number of times it appears in the research) and the indication of the average	CSV	topics-nodes.csv topics-links.csv

	Comfort and Energy levels which have been detected for this topic (for Emotional Analysis). The links file indicates the edges of the network graph, its links. Each row includes the labels for source and target nodes and the weight of the link (how many times it appears). Bidirectional links appear twice in the file, with the source and target exchanged.		

5.2.2. Topic Networks Datasets

These datasets represent the ways in which the principal topics described in the research (the ones featured in the focuses in the previous sections) relate with the other topics.

These datasets allow to study co-occurrences of topics, to discover possible correlations and, in general, to understand how discussions are formed, combining which topics (for example: if one talks frequently about Cybercrime and Cloud, it may be a good hint that the Cloud is one of the principal targets, in this person's opinion, of Cyberattacks).

These datasets come JSON format and are composed of two parts:

- **nodes**: they are the nodes of the network; each node represents a topic; each node has a label, with its name, and a weight, indicating how important is the topic (ie: how many times it appears in data); in the previous sections nodes with a higher weight correspond with larger circles
- **links**: they are the connections in the network, the edges of the graph; each link has a source and a target, indicating the label of the node from which the link comes from to the label of the node which the link goes to; if two nodes are linked bidirectionally, there will be two links, with A to B and with B to A in source and target; links also have a weight parameter,

indicating how important is that link (ie: how many times it appears); in the previous sections links with higher weight correspond with thicker connections between nodes

Title	**Description**	**Format**	**URL**
TagCloud_all	All the conversations with keyword frequency number	json csv	tagcloud_all_json.json tagcloud_all_csv.csv
TagRelations_links_all	All the relationships between the keywords	json csv	tagrelations_link_all_json.json tagrelations_link_all_csv.csv
TagRelations_nodes_all	All the nodes between the keywords	json csv	tagrelations_nodes_all_json.json tagrelations_nodes_all_csv.csv
Business_nodes	All the nodes between the Business keyword	json csv	business_nodes_json.json business_nodes_csv.csv
Business_links	All the relationships between the Business keyword	json csv	business_links_json.json business_links_csv.csv
Business_statistics	All statistics about the Business topic	json	Business_statistics_json.json Business_statistics_csv.csv

		csv	
Disrupt_nodes	All the nodes between the Disrupt keyword	json csv	Disrupt_nodes_json.json Disrupt_nodes_csv.csv
Disrupt_links	All the relationships between the Disrupt keyword	json csv	Disrupt_links_json.json Disrupt_links_csv.csv
Disrupt_statistics	All statistics about the Disrupt topic	json csv	Disrupt_statistics_json.json Disrupt_statistics_csv.csv
Economy_nodes	All the nodes between the Economy keyword	json csv	Economy_nodes_json.json Economy_nodes_csv.csv
Economy_links	All the relationships between the Economy keyword	json csv	Economy_links_json.json Economy_links_csv.csv
Economy_statistics	All statistics about the Economy topic	json csv	Economy_statistics_json.json Economy_statistics_csv.csv
Education_nodes	All the nodes between the Education	json	Education_nodes_json.json

	keyword	csv	Education_nodes_csv.csv
Education_links	All the relationships between the Education keyword	json csv	Education_links_json.json Education_links_csv.csv
Education_statistics	All statistics about the Education topic	json csv	Education_statistics_json.json Education_statistics_csv.csv
Industry_nodes	All the nodes between the Industry keyword	json csv	Industry_nodes_json.json Industry_nodes_csv.csv
Industry_links	All the relationships between the Industry keyword	json csv	Industry_links_json.json Industry_links_csv.csv
Industry_statistics	All statistics about the Industry topic	json csv	Industry_statistics_json.json Industry_statistics_csv.csv
Jobs_nodes	All the nodes between the Jobs keyword	json csv	Jobs_nodes_json.json Jobs_nodes_csv.csv
Jobs_links	All the relationships between the Jobs	json	Jobs_links_json.json

	keyword	csv	Jobs_links_csv.csv
Jobs_statistics	All statistics about the Jobs topic	json csv	Jobs_statistics_json.json Jobs_statistics_csv.csv
Manufacturing_nodes	All the nodes relationships between the Manufacturing keyword	json csv	Manufacturing_node_json.json business_nodes_csv.csv
Manufacturing_links	All the relationships between the Manufacturing keyword	json csv	Manufacturing_links_json.json Manufacturing_links_csv.csv
Manufacturing_statistics	All statistics about the Manufacturing topic	json csv	Manufacturing_statistics_json.json Manufacturing_statistics_csv.csv
Sharing_nodes	All the nodes between the Sharing keyword	json csv	Sharing_nodes_json.json Sharing_nodes_csv.csv
Sharing_links	All the relationships between the Sharing keyword	json csv	Sharing_links_json.json Sharing_links_csv.csv

Sharing_statistics	All statistics about the Sharing topic	json csv	Sharing_statistics_json.json Sharing_statistics_csv.csv
Skills_nodes	All the nodes between the Skills keyword	json csv	Skills_nodes_json.json Skills_nodes_csv.csv
Skills_links	All the relationships between the Skills keyword	json csv	Skills_links_json.json Skills_links_csv.csv
Skills_statistics	All statistics about the Skills topic	json csv	Skills_statistics_json.json Skills_statistics_csv.csv
Tech_nodes	All the nodes between the Tech keyword	json csv	Tech_nodes_json.json Tech_nodes_csv.csv
Tech_links	All the relationships between the Tech keyword	json csv	Tech_links_json.json Tech_links_csv.csv
Tech_statistics	All statistics about the Tech topic	json csv	Tech_statistics_json.json Tech_statistics_csv.csv

Work_nodes	All the nodes between the Work keyword	json csv	Work_nodes_json.json Work_nodes_csv.csv
Work_links	All the relationships between the Work keyword	json csv	Work_links_json.json Work_links_csv.csv
Work_statistics	All statistics about the Work topic	json csv	Work_statistics_json.json Work_statistics_csv.csv
NetNeutrality_nodes	All the nodes between the Net Neutrality keyword	json csv	NetNeutrality_nodes_json.json NetNeutrality_nodes_csv.csv
NetNeutrality_links	All the relationships between the NetNeutrality keyword	json csv	NetNeutrality_links_json.json NetNeutrality_links_csv.csv
NetNeutrality_statistics	All statistics about the NetNeutrality topic	json csv	NetNeutrality_statistics_json.json NetNeutrality_statistics_csv.csv
Privacy_nodes	All the nodes between the Privacy keyword	json csv	Privacyt_nodes_json.json Privacy_nodes_csv.csv

Privacy_links	All the relationships between the Privacy keyword	json csv	Privacy_links_json.json Privacy_links_csv.csv
Privacy_statistics	All statistics about the Privacy topic	json csv	Privacy_statistics_json.json Privacy_statistics_csv.csv
Security_nodes	All the nodes between the Security keyword	json csv	Security_nodes_json.json Security_nodes_csv.csv
Security_links	All the relationships between the Disrupt keyword	json csv	Security_links_json.json Security_links_csv.csv
Security_statistics	All statistics about the Security topic	json csv	Security_statistics_json.json Security_statistics_csv.csv
Cybercrime_nodes	All the nodes between the Cybercrime keyword	json csv	Cybercrime_nodes_json.json Cybercrime_nodes_csv.csv
Cybercrime_links	All the relationships between the Cybercrime keyword	json csv	Cybercrime_links_json.json Cybercrime_links_csv.csv

Cybercrime_statistics	All statistics about the Cybercrime topic	json csv	Cybercrime_statistics_json.json Cybercrime_statistics_csv.csv
Openness_nodes	All the nodes between the Openness keyword	json csv	Openness_nodes_json.json Openness_nodes_csv.csv
Openness_links	All the relationships between the Openness keyword	json csv	Openness_links_json.json Openness_links_csv.csv
Openness_statistics	All statistics about the Openness topic	json csv	Openness_statistics_json.json Openness_statistics_csv.csv
Democracy_nodes	All the nodes between the Democracy keyword	json csv	Democracy_nodes_json.json Democracy_nodes_csv.csv
Democracy_links	All the relationships between the Democracy keyword	json csv	Democracy_links_json.json Democracy_links_csv.csv
Democracy_statistics	All statistics about the Democracy topic	json csv	Democracy_statistics_json.json Democracy_statistics_csv.csv

Government_nodes	All the nodes between the Government keyword	json csv	Government_nodes_json.json Government_nodes_csv.csv
Government_links	All the relationships between the Government keyword	json csv	Government_links_json.json Government_links_csv.csv
Government_statistics	All statistics about the Government topic	json csv	Government_statistics_json.json Government_statistics_csv.csv
FakeNews_nodes	All the nodes between the FakeNews keyword	json csv	FakeNews_nodes_json.json FakeNews_nodes_csv.csv
FakeNews_links	All the relationships between the FakeNews keyword	json csv	FakeNews_links_json.json FakeNews_links_csv.csv
FakeNews_statistics	All statistics about the FakeNews topic	json csv	FakeNews_statistics_json.json FakeNews_statistics_csv.csv
IoE_nodes	All the nodes between the IoE keyword	json csv	IoE_links_json.json IoE_links_csv.csv

IoE_links	All the relationships between the Disrupt keyword	json csv	IoE_links_json.json IoE_links_csv.csv
IoE_statistics	All statistics about the IoE topic	json csv	IoE_statistics_json.json IoE_statistics_csv.csv
IoT_nodes	All the nodes between the IoT keyword	json csv	IoT_nodes_json.json IoT_nodes_csv.csv
IoT_links	All the relationships between the IoT keyword	json csv	IoT_links_json.json IoT_links_csv.csv
IoT_statistics	All statistics about the IoT topic	json csv	IoT_statistics_json.json IoT_statistics_csv.csv
AI_nodes	All the nodes between the AI keyword	json csv	AI_nodes_json.json AI_nodes_csv.csv
AI_links	All the relationships between the AI keyword	json csv	AI_links_json.json AI_links_csv.csv

AI_statistics	All statistics about the AI topic	json csv	AI_statistics_json.json AI_statistics_csv.csv
BigData_nodes	All the nodes between the BigData keyword	json csv	BigDatat_nodes_json.json BigData_nodes_csv.csv
BigData_links	All the relationships between the BigData keyword	json csv	BigData_links_json.json BigData_links_csv.csv
BigData_statistics	All statistics about the BigData topic	json csv	BigData_statistics_json.json BigData_statistics_csv.csv
VR_nodes	All the nodes between the VR keyword	json csv	VR_nodes_json.json VR_nodes_csv.csv
VR_links	All the relationships between the VR keyword	json csv	VR_links_json.json VR_links_csv.csv
VR_statistics	All statistics about the VR topic	json csv	VR_statistics_json.json VR_statistics_csv.csv

AR_nodes	All the nodes between the AR keyword	json csv	AR_nodes_json.json AR_nodes_csv.csv
AR_links	All the relationships between the AR keyword	json csv	AR_links_json.json AR_links_csv.csv
AR_statistics	All statistics about the AR topic	json csv	AR_statistics_json.json AR_statistics_csv.csv
Robots_nodes	All the nodes between the Robots keyword	json csv	Robots_nodes_json.json Robots_nodes_csv.csv
Robots_links	All the relationships between the Robots keyword	json csv	Robots_links_json.json Robots_links_csv.csv
Robots_statistics	All statistics about the Robots topic	json csv	Robots_statistics_json.json Robots_statistics_csv.csv
BlockChain_nodes	All the nodes between the BlockChain keyword	json csv	BlockChain_nodes_json.json BlockChain_nodes_csv.csv

BlockChain_links	All the relationships between the BlockChain keyword	json csv	BlockChain_links_json.json BlockChain_links_csv.csv
BlockChain_statistics	All statistics about the BlockChain topic	json csv	BlockChain_statistics_json.json BlockChain_statistics_csv.csv
Algorithms_nodes	All the nodes between the Algorithms keyword	json csv	Algorithms_nodes_json.json Algorithms_nodes_csv.csv
Algorithms_links	All the relationships between the Algorithms keyword	json csv	Algorithms_links_json.json Algorithms_links_csv.csv
Algorithms_statistics	All statistics about the Algorithms topic	json csv	Algorithms_statistics_json.json Algorithms_statistics_csv.csv
Predictive_nodes	All the nodes between the Predictive keyword	json csv	Predictive_nodes_json.json Predictive_nodes_csv.csv
Predictive_links	All the relationships between the Predictive keyword	json csv	Predictive_links_json.json Predictive_links_csv.csv

Predictive_statistics	All statistics about the Predictive topic	json csv	Predictive_statistics_json.json Predictive_statistics_csv.csv

5.3. Licensing and Contributions

All the open datasets coming from the research are made available under the **Open Data Commons Attribution License**.

The full legal text for the license can be found at the following link:

https://opendatacommons.org/licenses/by/1.0/

This licensing scheme means that

You are free:

- *To Share:* To copy, distribute and use the database.
- *To Create:* To produce works from the database.
- *To Adapt:* To modify, transform and build upon the database.

As long as you:

- *Attribute:* You must attribute any public use of the database, or works produced from the database, in the manner specified in the license. For any use or redistribution of the database, or works produced from it, you must make clear to others the license of the database and keep intact any notices on the original database.

Disclaimer:

This is not a license. It is simply a handy reference for understanding the ODC-BY 1.0 — it is a human-readable expression of some of its key terms. This document has no legal value, and its contents do not appear in the actual license. Read the full ODC-BY 1.0 license text for the exact terms that apply.

5.4. Contributions

While it is not required by the license, parties wishing to use the open data provided in this research are strongly encouraged to get in touch with us, to share your findings, visualizations, corrections, doubts, and to start collaborative projects which use these data and other we might collect and analyze together.

To get in touch, contact

nnexes


HER


PLUSVALUE