D6.3: Social media monitoring tools for assessment and support of engagement

WP6 – Engagement strategies for user participation

Document Information

Grant Agreement Number	688363	**Acronym**	hackAIR
Full Title	Collective awareness platform for outdoor air pollution		
Start Date	1st January 2016	**Duration**	36 months
Project URL	www.hackAIR.eu		
Deliverable	D 6.3 - Social media monitoring tools for assessment and support of engagement		
Work Package	WP6 – Engagement strategies for user participation		
Date of Delivery	**Contractual** 31st August 2017	**Actual**	31st August 2017
Nature	Demonstrator	**Dissemination Level**	Public
Lead Beneficiary	CERTH		
Responsible Author	Eleftherios Spyromitros-Xioufis (CERTH)		
Contributions from	Symeon Papadopoulos (CERTH), Stefanos Vrochidis (CERTH), Yiannis Kompatsiaris (CERTH), Christodoulos Keratidis (DRAXIS), Panagiota Syropoulou (DRAXIS), Carina Veeckman (VUB), Wiebke Herding (ONSUBJECT)		

Document History

Version	Issue Date	Stage	Description	Contributor
0.1	21/07/2017	Draft	Document structure	E. Spyromitros-Xioufis (CERTH), S. Papadopoulos (CERTH), S. Vrochidis (CERTH)
0.4	18/08/2017	Draft	Integrated document	E. Spyromitros-Xioufis (CERTH)
0.7	22/08/2017	Draft	Internal review of the whole document	S. Papadopoulos (CERTH), S. Vrochidis (CERTH), Y. Kompatsiaris (CERTH), C. Keratidis (DRAXIS), P. Syropoulou (DRAXIS)
0.9	29/08/2017	Pre-final	Final version reviewed by Carina Veeckman (VUB) and Wiebke Herding (ONSUBJECT)	S. Papadopoulos (CERTH), E. Spyromitros-Xioufis (CERTH), Carina Veeckman (VUB), Wiebke Herding (ONSUBJECT)
1.0	31/08/2017	Final	Approved by coordinator for submission	S. Papadopoulos (CERTH)

Disclaimer

Copyright message

Table of Contents

Table of Figures

1 Executive summary

This deliverable describes the social media monitoring tools that were developed within WP6 (T6.3) with the goal of supporting the implementation and assessment of the communication and engagement strategies of the hackAIR project, and especially activities in WP6 (Engagement strategies), WP7 (Pilot Support and Impact Assessment), and WP8 (Communication, Dissemination and Exploitation). Our work is based on an open-source software framework for monitoring, analysis and search over multiple social media platforms that we deployed, adapted and extended for the needs of hackAIR. The framework offers a powerful tool for hackAIR administrators and communication managers for assisting them in their social media audience management tasks: a) tracking social media discussions around keywords and user accounts of interest, b) obtaining real-time analytics over the tracked keywords/accounts, c) identifying new accounts and d) analyzing the structure of communities that are of relevance to hackAIR.

The deliverable focuses on two important extensions of the framework that were explicitly implemented to cover the needs of the hackAIR project with respect to: a) account discovery and b) audience analysis.

- The first aims at providing an automated hackAIR-related account discovery mechanism to the users of the tool and minimizing the amount of manual effort required to perform the discovery. This is accomplished by employing machine learning techniques to automatically score the discovered accounts with respect to their relevance with air quality topics and by exploiting the network structure of Twitter to find additional relevant accounts.
- The second extension aims at enhancing the audience analysis capabilities of the tool. In particular, we employ network analysis algorithms to discover communities within the audience of hackAIR accounts, as well as nodes (accounts) that are central and popular within each community and which could be possibly identified and engaged as hackAIR ambassadors. Using the hackAIR follower social network makes it also possible to better quantify the impact of hackAIR social media posts and tune the communication strategy.

At the time of delivery of this report, the described social media monitoring tool and its extensions have been made available to the hackAIR consortium and tests of its capabilities have been carried out by consortium partners, and have already led to interesting insights and valuable data. In addition, a first round of valuable feedback has been already received with respect to potential improvements and extensions of the tool. In the coming period, the tool will be used, tested and continuously improved in the context of supporting the engagement and communication activities of the project.

2 Introduction

Social media is nowadays a key channel for communication, self-expression, information gathering and sharing for billions of Internet users worldwide. The penetration of social media has consistently increased throughout the last decade and it has nowadays reached a point where the large majority of people worldwide regularly use one or more of numerous popular platforms for a wide variety of purposes. As a result, social media is often regarded as a sensor of real-world trends and events (Aiello, et al., 2013) and as an indispensable tool for performing social science research (Mejova, Weber, & Macy, 2015) and consumer intelligence gathering and marketing (Fan & Gordon, 2014).

Given the increasing importance of social media for information sharing and intelligence gathering, the hackAIR work plan has foreseen work on social media monitoring in order to support and reinforce its key research and innovation activities. In particular, the deployment of advanced social media monitoring technologies is seen as valuable for the following project activities:

- **Engagement strategy to encourage involvement** (T6.1): The engagement strategy (documented in D6.1) advises pilot coordinators and local communication managers to make use of social media monitoring tools to support engagement-oriented activities during pilot trials, and especially during the full pilot implementation phase. With the help of social media monitoring, pilot coordinators and local communication managers can identify hackAIR "ambassadors", and possibly establish connections with them. hackAIR ambassadors are influential users in the domain of air quality, who can help in finding new participants for the project and in reaching out to a wider audience. Ambassadors are leading communicators on social media, and therefore, it is advised that pilot coordinators and local communication managers, with the support of CERTH, regularly use the outputs of social media monitoring tools.
- **Pilot support and impact assessment** (T7.3 & T7.4): Being able to monitor different social media streams and discover users in different locations, who are active in air quality issues, will make it possible to get in touch, engage and attract additional citizens to the pilot activities. Moreover, the capabilities of the presented social media monitoring tools will support the extraction of impact indicators, and more specifically the impact and reception that the hackAIR messages and results have on the different social media communities.
- **Communication and dissemination strategy and content** (T8.1 & T8.2): Communication and dissemination can greatly benefit from the capabilities of the delivered monitoring tools. First, the tools can help communication managers discover relevant content from a diverse set of sources that can be then used as part of hackAIR's social media posting activities, making the hackAIR accounts more attractive and interesting. Second, the tools make it possible to quantify the reach of hackAIR's posts across different social media communities and to identify influencers who can then be engaged by the communication managers with the goal of contributing to the wider dissemination of hackAIR results. Finally, the audience analysis capabilities of the tools can lead to valuable insights that can be used by social media managers to fine-tune their communication strategy.

To this end, an existing **open-source social media monitoring framework** was configured and deployed to serve as the technical basis for the project. This framework is a stand-alone software system that is installed, managed and operated independently of the hackAIR platform. Technical details on the design, implementation and usage of this framework are provided in Section 3, while it is worth noting that the work was accepted for publication in the 4th Internet Science Conference[1] (INSCI 2017) as a technical demonstration paper (Schinas, Papadopoulos, Apostolidis, Kompatsiaris, & Mitkas, 2017). At the time of delivering this report, the tool is operational and available on the following link: http://hackair-mklab.iti.gr[2].

A first extension to the framework of Section 3 was the development of a sophisticated methodology for **discovering hackAIR-relevant accounts** on different social media channels. The methodology, which is described in Section 4, is based on the collection of a large initial set of candidate user accounts from different social media platforms, and then on the application of automatic classifiers to separate between relevant and irrelevant accounts. The collection

[1] 4th International Conference on Internet Science, Nov 22-24, 2017, Thessaloniki, Greece: http://internetscienceconference.eu/

[2] The page is password protected to prevent unauthorized usage by third parties. The credentials are available upon request by the main authors, Eleftherios Spyromitros-Xioufis (espyromi@iti.gr) and Symeon Papadopoulos (papadop@iti.gr).

of candidate accounts was based on querying the respective platform APIs with a list of air quality-related keywords (that was compiled with the help of several hackAIR partners), while additional candidate accounts were discovered with the help of an iterative exploration process based on Twitter lists and by monitoring air quality-related streams of tweets. The classification of accounts into relevant and irrelevant was based on classification models that were constructed based on text features extracted from the accounts' profile metadata (description) and their posts. Many of the discovered accounts (indicative examples are presented in Tables Table 7-Table 8) were found by hackAIR partners as relevant for being engaged in the project pilot and communication activities.

A second extension to the social media monitoring framework of Section 3 concerned the development of a solid methodology for **analyzing the social media audience of hackAIR**. Moving beyond the existing analytics tools of social media platforms, this methodology leverages social network analysis tools to create a more structured view of the hackAIR social media audience. Specifically, the network of hackAIR followers and their followers is analyzed with the help of a community detection algorithm to extract groups of Twitter accounts that are more densely connected to each other, and therefore correspond to topically and geographically focused communities. This result combined with the analysis of network connectivity of Twitter accounts led us to the discovery of influencers, i.e. accounts that are followed by many other accounts and that have an important role in their community. Several examples of such accounts, both within the hackAIR follower base (Table 13) and outside of it (Table 14), are considered important by the hackAIR communication managers for fine tuning their communication strategy and activities.

Concluding this report, Section 6 provides an outline of future work, and more specifically a number of ways that the delivered results will be actively utilized in the context of pilot engagement, communication and dissemination activities.

A note on privacy: Similar to many research and innovation activities that involve data collection from social media sources, the work presented here could potentially raise privacy issues, e.g. the tool may inadvertently collect posts that reveal personal information about the users that published them. During the development of the tool, we have carefully considered such issues and have taken reasonable measures to make sure that the privacy rights of social media users are in no way put at risk. As will become obvious from the following sections, a number of classifiers have been developed to separate between hackAIR-relevant and hackAIR-irrelevant content (the latter of which is not stored and processed in any other by our tools). Manual inspection of numerous hackAIR-relevant posts has shown that they are of public nature, i.e. they do not reveal personal information, but rather pertain to news items and announcements related to air quality topics (which are meant to be publicized). In addition, an extension of the tool is currently implemented to automatically delete raw data at a fixed short time period (configurable, default value of one month) after their collection. Finally, it is worth noting that DRAXIS submitted a formal notification to the Greek Data Protection Authority (on July 2017), informing about the data collection activities of the project that also include the data collection activities described in this report.

3 Social media monitoring technology

The hackAIR social media monitoring solution is based on a framework that has been developed by CERTH with the goal of addressing the limitations and shortcomings of the two main types of existing solutions that are presently used for social media monitoring:

- In academic settings, researchers typically use a variety of tools, scripts or libraries, and often develop their *own additional custom scripts* to perform the required data collection, manipulation and analysis.
- In business settings, analysts use a variety of *software-as-a-service products*, typically through a dashboard-like interface, offering access to statistics and analytics about queries of interest.

The main disadvantage of the first approach is the increased effort that is necessary to set up the data collection and analysis pipeline. In contrast, the second approach suffers from limited flexibility, cost (most social media SaaS offerings are subscription-based) and dependence on proprietary solutions.

To address these limitations, we based the hackAIR social media monitoring solution on an integrated open-source framework that has been developed by CERTH. In particular, the development of the framework started in the FP7 SocialSensor project (http://socialsensor.eu) and continued in the FP7 REVEAL project (https://revealproject.eu), while the currently presented version is based on extensions made to explicitly support the social media monitoring needs of the hackAIR and STEP (http://step4youth.eu) projects, both of which were in need of an easily deployable solution with a user-friendly dashboard and a simple REST API. As a result, the current version of the framework is easy to install and manage, and offers a web-based user interface for setting up tracking tasks and browsing the collected data and extracted analytics. Importantly, the framework is designed in a modular way, making it straightforward to adapt and extend according to different requirements (e.g. collecting data from additional social media platforms, implementing additional content filters, computing additional metrics, etc.). As a result, in addition to hackAIR, the tool could be a convenient choice for several other projects (especially CAPS) that are in need of monitoring discussions and engaging with social media communities.

3.1 Design and implementation details

The presented tool is based on a set of publicly available GitHub projects[3]. The core project is *mklab-framework-common* and contains classes for defining a common data model for several social media platforms, while m*klab-socialmedia-abstractions* implements the data models for each of the supported platforms and contains wrappers that encapsulate the different APIs that are used for the collection of data. The *mklab-framework-client* project contains classes for data management, retrieval and analytics, such as wrappers around different storage solutions such as MongoDB and Apache Solr. Finally, the *mklab-stream-manager* project implements the orchestration and operational logic for data collection and storage.

3.1.1 Data model

Figure 1 depicts the underlying data model. In the left side of the diagram, we see the data objects collected by the tool: items, media items, web pages and users. Items represent the messages posted to the various social media platforms, e.g., status updates on Twitter, posts on Facebook, videos on YouTube, etc. Media items correspond to the multimedia content that is embedded in these items, for instance, videos in tweets, images in Facebook posts, etc. Web pages correspond to the URLs contained in some items. Finally, user entities correspond to the accounts that are publishing the items. The right side of the diagram depicts the collection entity, which is the construct used to organize social media content collected by the tool. A collection comprises multiple queries as described in the next section, while the same query can be included in more than one collection. For hackAIR, we use two query types for data collection: a) keyword queries, and b) account queries[4].

[3] All of these projects are hosted by the MKLab-ITI GitHub organization: https://github.com/MKLab-ITI/

[4] Although the tool also supports location queries (i.e. collection of content that is geotagged within a given bounding box), we found that such queries were very restrictive and led to very limited results due to the small percentage of geotagged content in social media. For that reason we focused on the two other query types.

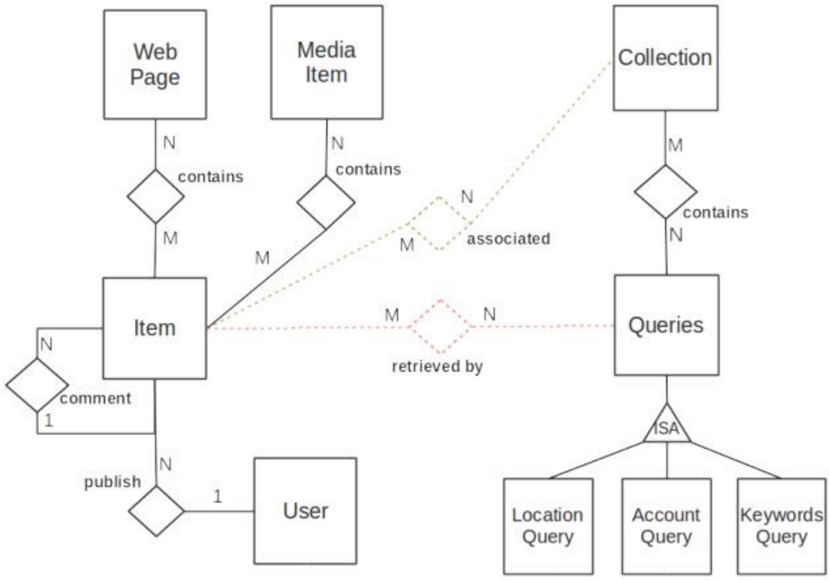

Figure 1: Entity relationship data model of social media monitoring tool

3.1.2 Data collection and processing

The collection of data from the target social media platforms is performed by the *Stream Manager* module. This collects posts from Twitter, Facebook, YouTube and Google+, alongside the user accounts that published them and the embedded media items and linked URLs. In addition to social media sources, the platform also supports monitoring RSS feeds. Data collection by the Stream Manager is based on user-defined collections, which consist of a set of queries of the following two types: keyword-based and account-based. Keyword-based queries may consist of simple keywords, e.g. "*air pollution*", or complex logical expressions, e.g. "*haze AND (pm10 OR pm2.5)*". User account-based queries refer to *public* sources of content for each platform, e.g., Twitter users, Facebook pages, etc.

Given a collection, the associated queries are translated to appropriate API calls. For example, if a Twitter account such as @EEB (the official account of the European Environmental Bureau) is specified as one of the queries in the collection, that account is mapped to a call to the Twitter API that is used for the collection of Tweets posted by it. In the same way, given a keyword or a logical expression of keywords, multiple API calls are generated, one for each of the supported sources. However, given that some of the defined collections may contain identical queries, e.g. the same keywords, the Stream Manager first de-duplicates all input queries into non-redundant basic query elements and then periodically polls each of the unique query elements, and keeps track of the number of submitted requests to each platform in order to respect the limits imposed by it. The fetched items are then processed by a sequence of filters and processors, executed within the Stream Manager, before being stored and indexed. In terms of filtering, a set of heuristic rules are applied to keep only items of high quality. For example, items with limited text content or with too many hashtags and URLs are treated as spam messages and therefore discarded. In terms of processing modules, three indicative processors are provided off-the-shelf by the Stream Manager: language detection, named entity extraction, and MinHash signature extraction from text (used for duplicate removal). Finally, the collected content elements (items, media items, users and web pages) are stored in a MongoDB[5] instance.

3.1.3 Indexing and retrieval

Indexing of the data is based on Apache Solr[6]. For each item, only those fields that can be used for filtering or faceting are indexed, e.g. all textual fields, publication time, user id, number of views, etc. Retrieval of content related to a specific collection is performed in two stages: a) first, via the generation of a Solr query that returns the ids of the relevant items, and b) second, via a retrieval-by-id operation that fetches the corresponding item metadata and all associated entities from MongoDB.

The tool offers several analytics capabilities that range from simple metrics such as the number of items or unique users, to more complex ones such as *reach* (estimated cumulative audience for a set of items) and *endorsement*

[5] https://www.mongodb.com/
[6] http://lucene.apache.org/solr/

(estimated sum of “likes" for a set of items). Top users, locations, tags and named entities (based on frequency of appearance) are also provided. Furthermore, the tool supports the generation and visualization of timelines with different time granularities. All these analytics operations are implemented on top of two Solr components: *Faceting* and *Stats*.

Finally, the analytics framework leverages two more Solr components to make easier the exploration of collections: Result Clustering and Result Collapse. Clustering is used to automatically discover groups of similar search hits, i.e. groups of items related to a specific topic. These identified topics can be used to narrow down the set of items associated with a collection. Collapse, on the other hand, is used to group search results based on the value of a field. In our case, we use collapse based on the MinHash signature of items to support de-duplication of content, i.e. items with the same signature collapse into a single item. Figure 2 depicts the key software components of the framework along with the function that each performs.

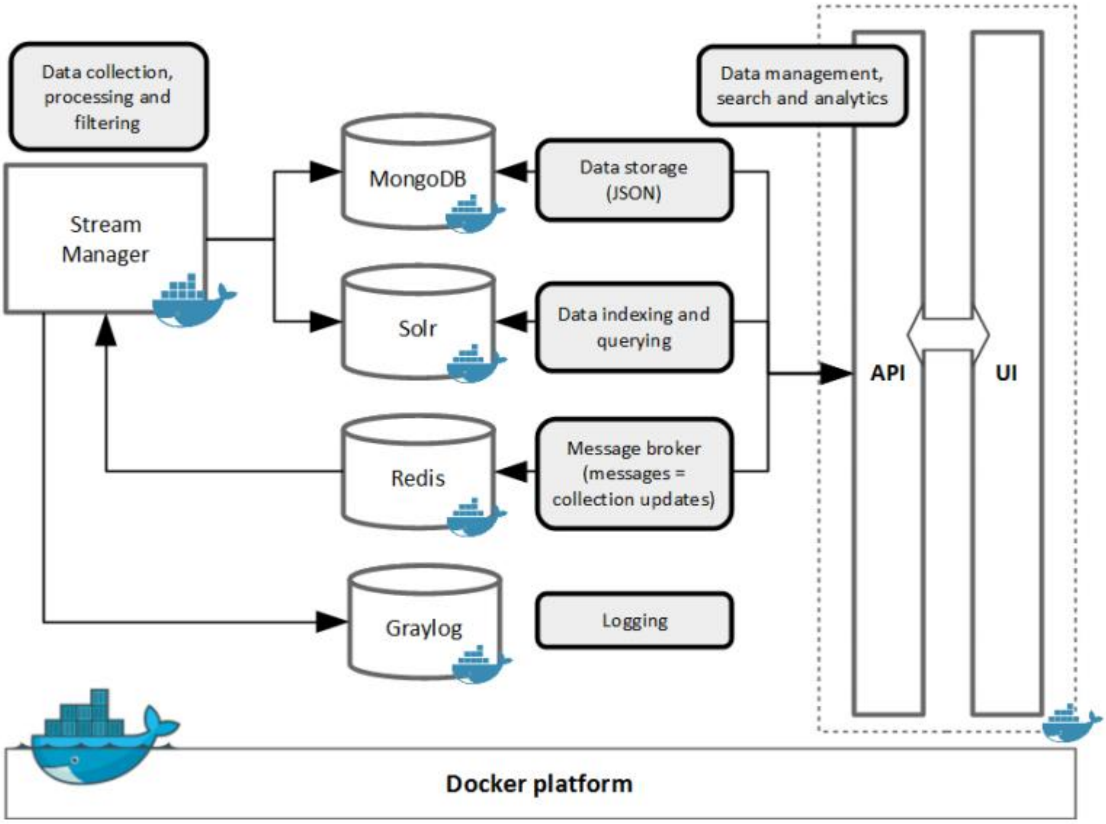

Figure 2: Overview of framework components (white blocks) and their function (gray).

3.2 Usage and user interface

In this section, we briefly describe the user interface and usage of the tool (step-to-step guidelines about the usage of the tool are available through a set of online tutorials[7] that were created for the needs of the STEP project). The user interface of the tool consists of three main pages: a) a page where users can define collections of media items (Figure 3), b) a *feed view* for browsing the content collected around a specific collection (Figure 4), and c) a *dashboard view* for analytics (Figure 5).

The collection definition page allows the user to specify the name of the collection and to add the accounts and/or the keywords (simple or complex) that he/she wants to be monitored. In the account specification field, the user can select the social media platform that should be searched and account suggestions are provided (in the form of auto-complete suggestions) to assist the user in the discovery task.

[7] https://www.youtube.com/playlist?list=PLA8q-g6RNMeACeprZYpEFr8Fw7bixeWWl

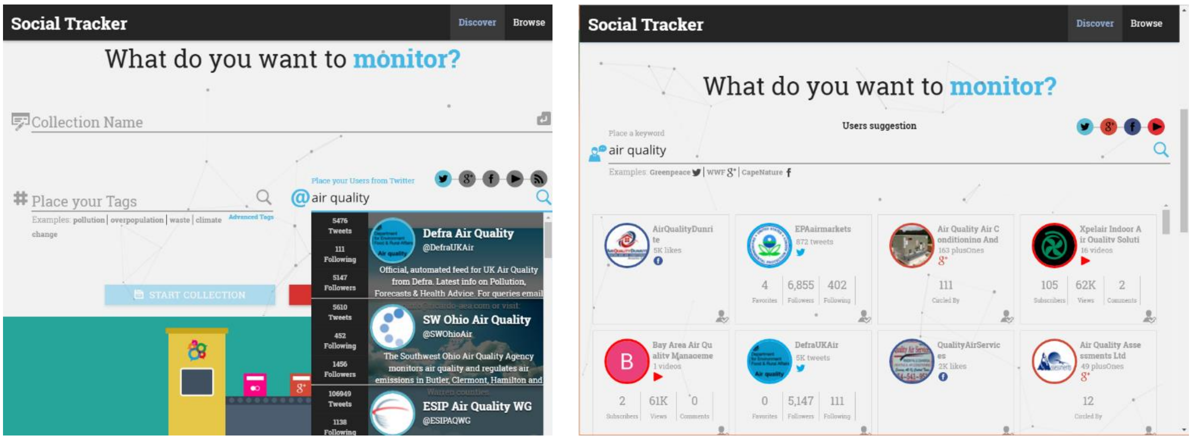

Figure 3: Collection definition page

The feed view presents the list of media items of the chosen collection in the form of a stream. For each item in the view relevant metadata are shown, namely publication time, username and social media platform. The feed can be presented in both gallery/grid (Figure 4) and list form. The user can also search for items using free text queries and use several filtering and sorting criteria. The following filters are provided: a) source (social media platform), b) language, c) topic, d) original (show retweets/shares or not), e) type (text item or item with embedded multimedia), f) unique (remove near-duplicate content), g) date.

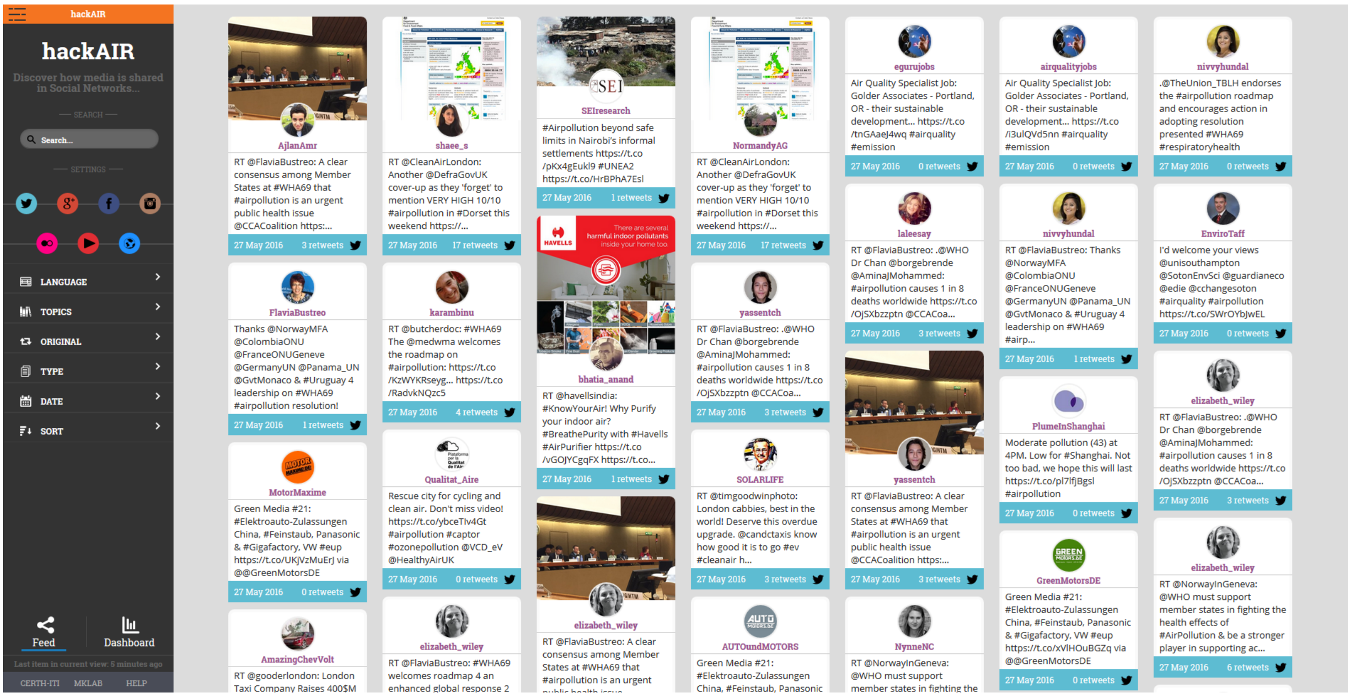

Figure 4: Feed view of the social media monitoring tool

The dashboard view of the tool offers metrics and widgets for generating summary views over the collected data. The visualizations are dynamic and can be updated by leveraging the same set of filters as the ones in the feed view. In more detail, the dashboard contains the following widgets:

- Numbers of posts made, users talking, users reached and endorsements, and platform contribution pie chart.
- Heatmap based on the exact location (latitude, longitude) of geo-tagged items in the collection (which typically correspond to a small subset of all items).
- User location at country level, automatically detected from the location text field in users' profiles.
- Timeline visualization based on the number of items over time with customizable granularity of depiction.
- Top N users with most posts in the collection. N can be set between 10 and 200. Avatar, username and total number of posts for each user is presented along with the link to their social profile pages.
- Top N entities in terms of frequency. N can be set between 10 and 200. Entities are organized in three frequency categories (often, occasionally, seldom) and three types (person, tag, location).

Figure 5: Dashboard view of the social media monitoring tool

The tool has been already tested by hackAIR partners and several collections have been created (8 at the time of writing this report). These include, for instance, topics such as *air quality and citizen observatories*, *air pollution and asthma*, *air pollution* in Brussels, *air pollution* using German keywords (*luftqualität, feinstaub*), air pollution using Norwegian keywords (*luft, luftforurensing, luftforurensning, luftkvalitet, svevestøv, dieselforbudt, luftforureining, måling luft*), and *PM10*.

3.3 Limitations

There are two main limitations with respect to the usage of the tool. The first is a result of the API usage restrictions imposed by social media platforms[8]. Restrictions to the number of API calls allowed in a fixed time window result in delays in the discovery of new content, especially for queries (keywords) that exhibit high activity, because it is necessary to decrease the request rate for each query. The second limitation pertains to the scalability of the tool. As the number of collected items reaches the order of tens of millions, the response time of the tool is adversely affected by the size of Solr index. This is more pronounced in the case of analytics, which are calculated on the fly using the faceting mechanism of Solr. Since the queries required for the social media monitoring needs of hackAIR are quite specific and do not involve a very large number of media items, these limitations do not have a noticeable impact on the hackAIR instance of the tool.

In terms of functionality, on the other hand, we found that there were two directions for extension of the tool so that it better serves the goals of hackAIR: a) developing a more sophisticated, automated mechanism for discovering relevant accounts and b) improving the audience analysis capabilities of the tool. Our work towards these two directions is presented in sections 4 and 5, respectively.

[8] We are referring to the free API versions.

4 Discovery of relevant social media accounts

Despite its merits, the initial version of the social media monitoring tool still required considerable effort in order for the user to perform advanced search queries and to distinguish between relevant and irrelevant results. To this end, we developed a sophisticated account discovery module that requires only the provision of an initial set of hackAIR-related keywords and then automatically generates and continuously refines a list of accounts of interest to hackAIR. As *hackAIR-relevant accounts*, we define social media accounts that regularly post about air quality/pollution issues. In addition, the module generates a number of statistics with respect to the activity, popularity and geographical distribution of the discovered accounts. The module currently runs as an independent background process, but its integration with the front-end of the social media monitoring tool is planned for the coming period.

4.1 Methodology

In this section, we describe in detail the methodology that we developed for the systematic discovery of hackAIR-relevant accounts. The methodology comprises several steps that are graphically illustrated in Figure 6. The first two steps deal with the creation of a seed list of multi-lingual keywords related to the topic of interest (in this case air quality) and their use as query terms in the APIs of social media platforms. In these steps, hackAIR partners (DRAXIS, BUND, NILU, and ONSUBJECT) provided feedback on the selected keywords and suggested additional ones that could be relevant to the topic, so that a set of candidate accounts can be generated by querying the APIs of multiple social media platforms (section 4.2). Then, the initial set of potentially relevant accounts is further expanded by employing a method that exploits account co-occurrence in Twitter lists to identify Twitter accounts that are similar with those retrieved in the first step and thus of potential relevance to hackAIR (section 4.3).

After a set of candidate accounts is generated in the previous steps, a filtering step (section 4.4) takes place that aims at discarding accounts that are not relevant with hackAIR despite having been retrieved with air quality-related keywords (we observed that this is commonly the case). To this end, a classifier is employed to classify accounts as relevant/irrelevant with air quality based on the text contained in the name and description fields of the accounts. To build the classifier, we used a set of training accounts that have been manually annotated as relevant/irrelevant with the help of hackAIR partners. The classifier will be regularly updated as additional feedback (training examples) from partners becomes available. The output of this step is a set of accounts that are very likely to be relevant with air quality topics. However, some of these accounts might be inactive for a long period and therefore not useful for our purposes. Thus, an additional filtering step is applied to discard such accounts (section 4.6).

A limitation of the account discovery pipeline that has been described so far is that it considers only accounts that contain air quality-related terms in their name or description fields while ignoring accounts that often post about air quality issues without having air quality-related terms in their profile (name and description). To this end, an additional step was realized (section 4.5) with the aim of detecting accounts that frequently post air quality-related content. This is achieved by collecting posts (Tweets in particular) that contain air quality-related terms over a long period of time and identifying accounts that consistently post about air quality. This set of accounts is combined with the accounts discovered based on their description and name fields to generate the final list of interesting accounts. As a final step, several statistics are calculated about these accounts such as popularity, activity and geographical distribution (section 4.7).

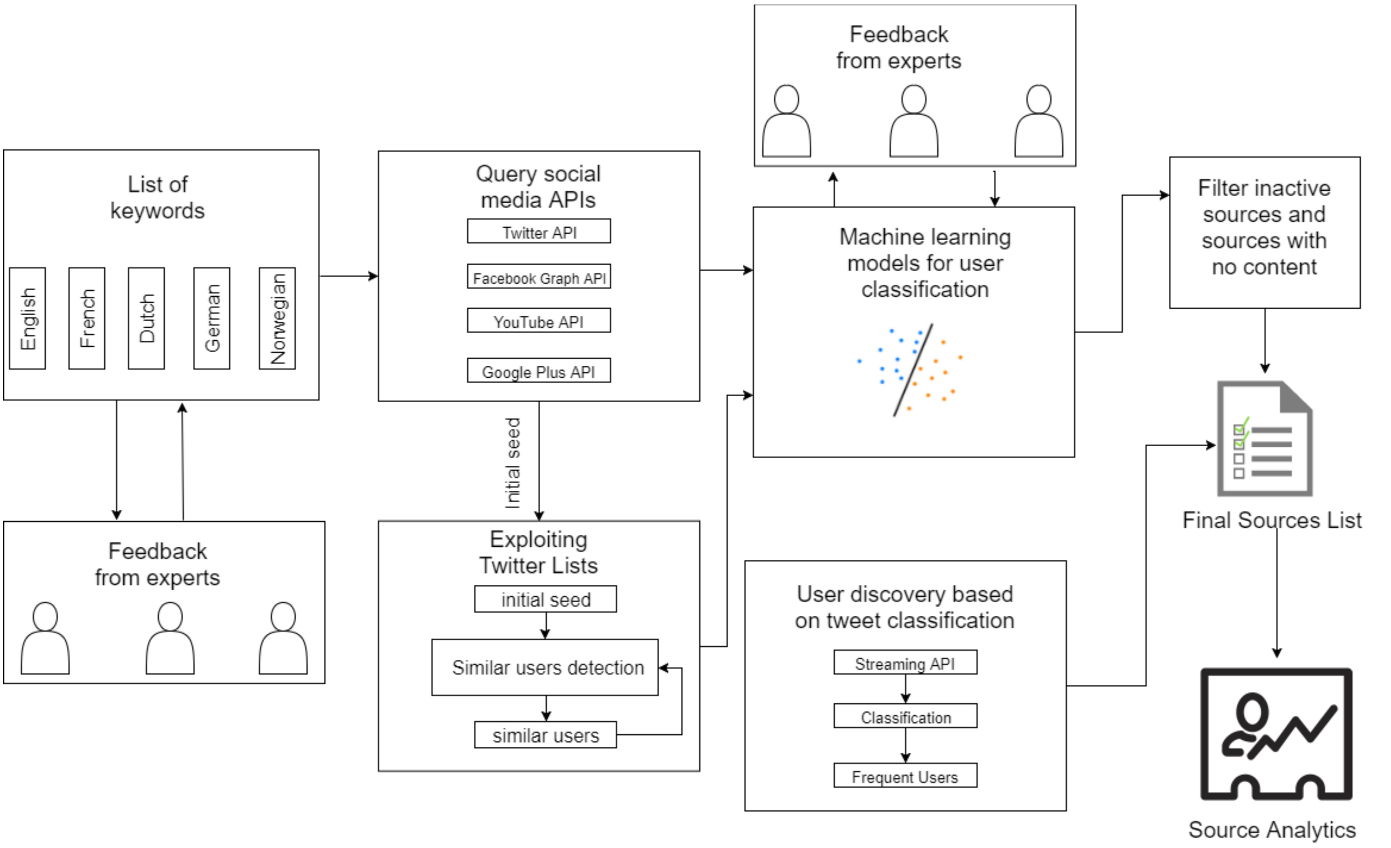

Figure 6: A schematic representation of the account discovery methodology

4.2 Keyword selection and querying social media APIs

The selection of appropriate keyword terms for querying the APIs of social media platforms plays a crucial role in the overall effectiveness of the account discovery module as it determines the initial pool of candidate accounts. To generate the list of keywords we worked closely with all project partners and especially with the engagement and communication managers (ONSUBJECT). As a result of this collaborative effort, a list of 527 multi-lingual (113 English, 102 German, 99 Norwegian, 94 French and 119 Dutch) air quality-related keywords was composed (e.g. *air pollution*, *luftforurensing*, *luftverschmutzung*, *la pollution de l'air*, *luchtvervuiling*)[9].

Having composed the list of search keywords, the next step of the pipeline is the retrieval of candidate accounts by querying the APIs of social media platforms. In particular, we used the APIs of Twitter, Facebook, YouTube and Google+. Note that we refer to all returned objects as *accounts* even though they are different in nature (e.g. Twitter and Google+ users, Facebook pages, YouTube channels). The following subsections provide the details of how each API was queried. The collected results are then processed by the subsequent steps of the pipeline of Figure 6.

4.2.1 Twitter

The Twitter REST API[10] provides programmatic access to read and write operations such as creating new Tweets, reading user profile and follower data, etc. In this case we use it to retrieve Twitter account information in JSON format using keywords. Specifically, we use the *users/search*[11] endpoint that provides a simple, relevance-based search interface to public Twitter accounts. Documentation of the API also states that only the first 1,000 matching results are available per keyword query. Information about each retrieved account includes the account's name, location, language, description, number of followers, number of tweets, etc. Querying this endpoint with the list of air quality-related keywords returns about 2.5K Twitter accounts.

[9] The full list of keywords used can be found in Table 12 of the Appendix.
[10] https://dev.twitter.com/rest/public
[11] https://dev.twitter.com/rest/reference/get/users/search

4.2.2 Facebook

In order to search for Facebook accounts, we use the Facebook Graph API[12], which is a low-level API that can be used to programmatically query data, post new stories, manage ads, upload photos, and perform a variety of other tasks that an app might implement. The Graph API supports search over many public objects in the social graph with the *search*[13] endpoint. The *search* endpoint can retrieve objects like Users, Pages, Places, Events and other. Since, User objects in Facebook typically represent actual persons and Graph API searches are based on object names, it would not make sense to search for User objects using the air quality-related keyword list. Facebook Pages, on the other hand, can represent organizations or causes and may therefore contain air quality-related terms as part of their names. Hence, we query the Graph API for Facebook pages containing air quality-related keywords. Information about each retrieved Page includes name, location, language, description, number of likes, etc. Querying the Graph API with the air quality-related keywords returns about 5.3K Facebook Pages.

4.2.3 YouTube

Additionally, we search for YouTube channels that can be relevant with our air quality issues. The YouTube Data API[14] is used for that purpose. There are different types of resources that can be retrieved using the API, as well as methods to insert, update, or delete resources. To search for YouTube channels with air quality-related keywords, we use the *search/list* endpoint[15] which returns a collection of search results that match the query parameters specified in the API request. By default, the set of results contains all matching video, channel and playlist resources but can be configured to return only YouTube channels. For each channel, information such as name, description, number of subscribers/videos/views, etc. is provided. Querying the YouTube Data API with air quality-related keywords, we collect about 2.5K YouTube channels.

4.2.4 Google+

Finally, we attempt to discover air quality-related accounts from Google+ by using the respective API[16]. Similarly to Facebook, Google+ contains various types of object such as Users, Pages, etc. Unfortunately, the part of the API that provides access to Pages is not publicly available (special restrictions apply) and we therefore focus only on Google+ Users. To search for all Users with public profiles using air quality-related keywords we use the *People:search* method[17]. Information about each retrieved user includes name, description, number of followers (circledBy), number of +1s, etc. Querying the Google+ API with air quality-related keywords resulted in the collection of about 350 Google+ users.

4.3 Exploiting Twitter lists

We try to further extend the set of candidate accounts by exploiting the concept of Twitter Lists[18] and leveraging the method presented in (Kanungsukkasem & Leelanupab, 2016). A Twitter List is a group of Twitter accounts that allows users to see tweets from a group of people in individual tweet timelines, without having to put them all together into their own timeline of people they follow. Twitter users can create their own lists or subscribe to lists created by others. Thus, one may expect that accounts appearing in the same list will likely post similar content and will therefore be similar to each other. Based on this observation, Kanungsukkasem & Leelanupab (2016) proposed an effective and lightweight algorithm of finding similar users.

[12] https://developers.facebook.com/docs/graph-api
[13] https://developers.facebook.com/docs/graph-api/using-graph-api/
[14] https://developers.google.com/youtube/v3/docs/
[15] https://developers.google.com/youtube/v3/docs/search/list
[16] https://developers.google.com/+/web/api/rest/
[17] https://developers.google.com/+/web/api/rest/latest/people/search
[18] https://dev.twitter.com/rest/reference/get/lists/list

Our implementation of this algorithm works as follows: Let T be the target account for which we want to find similar accounts. In the first step, we retrieve all Twitter lists L_i in which T is a member using the *lists/memberships*[19] endpoint of the Twitter API. Then, the *lists/members* endpoint[20] is used to retrieve all other members (accounts) in the lists. Let U denote the set of unique accounts that were retrieved from all L_i Lists. The similarity of each account $u \in U$ with the target account T is computed by counting the number of times u appears in the same list with T, and then normalizing by dividing with N, the number of lists that the target account belongs to. To tell whether two accounts are similar enough, a similarity threshold $\boldsymbol{\vartheta}$ must be defined. Taking into account the fact that the account discovery pipeline includes a sophisticated filtering step that will remove most of the irrelevant accounts, we opt for using a low threshold ($\boldsymbol{\vartheta}$ =0.3), in order to allow for more potentially relevant accounts to be discovered.

The algorithm described above is first executed using as seeds 10 Twitter accounts that are manually selected to be related to air pollution and also highly popular and active. This first execution resulted in approximately 170 similar accounts that were then used as targets in a second execution of the algorithm that returned approximately 1077 new similar accounts. Using more seed accounts in the first execution or performing more iterations of the algorithm was not possible in reasonable time due to call rate limitations of the Twitter API.

4.4 Account classification

The previous steps of the account discovery pipeline led to a collection of around **12K accounts** in total that contain air quality-related terms in their name and description fields or are similar to such accounts. However, we found that a significant portion of these accounts are not relevant for hackAIR, e.g. because they concern more general environmental issues. Table 1 provides some indicative examples of relevant and irrelevant accounts. As a result, it was considered necessary to create an intelligent filtering mechanism for detecting and discarding irrelevant accounts. To this end, we employed machine learning: we trained a classifier to distinguish between relevant and irrelevant accounts based on textual metadata such as their name and description.

Table 1: Indicative examples of relevant and irrelevant accounts

Name	description	label	reason
Birdi	More than a smoke detector. The Birdi Smart Detector tracks air quality, whether it's a health danger like indoor air pollution or an emergency like fire & CO.	Irrelevant	Commercial smoke detector product
Breathing London	Monitoring the Air Pollution in London since 2010. - data from http://t.co/YK6JzIRq	Relevant	Continuous air quality information
Drive Electric RVA	A chapter of @EAANational, we promote electric vehicles as a fun, practical transportation alternative that also reduces energy consumption and air pollution.	Irrelevant	Electric car news
BeijingAir	MetOne BAM 1020 and Ecotech EC9810 monitors, reporting PM2.5 and ozone readings. Format for each: pollutant type; concentration; AQI; definition.	Relevant	Real time air quality monitoring for Beijing
Air Pollution Music Pty Ltd	Mixing Business with pleasure	Irrelevant	Music related source
Plume Labs	Proud builders of Flow, the first smart mobile air quality tracker, and the Plume Air Report, the free app to track live air pollution forecasts around the world.	Relevant	Tracking live air pollution forecasts
TheMnPCA	Working with Minnesotans to protect and improve the environment and enhance human health. The Minnesota Pollution Control Agency (MPCA) monitors environmental quality, offers technical and financial assistance, and enforces environmental regulations.	Irrelevant	General environmental source

[19] https://dev.twitter.com/rest/reference/get/lists/memberships
[20] https://dev.twitter.com/rest/reference/get/lists/members

To build the classifier (*Air_Quality_Account*) we created a training set of 1000 examples that were manually labeled as relevant or irrelevant by reading the textual fields of the account profile (name and description). The text was preprocessed by applying stemming and stop-word removal and represented using a tf-idf bag-of-words representation. As classification algorithm, we used L2-regularized L2-loss Support Vector Machines (the LibLinear[21] implementation) with default parameters. Note that the classifier was trained using accounts with textual metadata in different languages and is therefore applicable to all accounts of the collection.

Since the accuracy of the classifier depends heavily on the size and the quality of the annotations in the training set, to further improve the classifier we asked project partners to help us improve the composition of the training set. In particular, we provided them with a large list of accounts along with the classifier's decisions and asked them to manually check the correctness of each classification. This way, a new training set of 1150 examples was created. Table 2 provides details about the *Air_Quality_Account* classifier before and after feedback. We notice that after feedback the training set was slightly expanded and that classifier performance improved (measured by Precision and Recall that were estimated using 10-fold cross-validation). After applying the refined classifier to the whole collection of 12K candidate accounts, 1388 accounts were classified as relevant.

Table 2: Air_Quality_Account classifier details

Classifier	# examples	# relevant	# irrelevant	Precision	Recall
Before feedback	1000	400	600	85.0%	85.4%
After feedback	1150	450	700	87.6%	86.8%

4.5 Account discovery based on tweet classification

In the previous sections, we focused on discovering accounts that contain air quality-related text in their name or description. Here, we describe an approach for detecting Twitter accounts that post air quality-related information but were not possible to retrieve with the previously discussed methods. In short, the approach consists of collecting Tweets from specific locations for a long period of time, using a classifier similar to the one presented above to detect Tweets that are related to air quality and identifying accounts that regularly post Tweets of this type.

Here, we present our experiments to evaluate the approach. For these experiments, we used a large dataset of Tweets that was initially collected to facilitate Twitter-based air quality estimation experiments that were presented in D3.2. The dataset contains approximately 1M Tweets that contain air quality-related terms and were collected over a period of about seven months (1/2/2017 until 20/7/2017) around five cities in the UK (London, Liverpool, Manchester, Birmingham and Leeds). Details about how this dataset was collected can be found in section 4.2.1 of D3.2; the following paragraph provides details about how Tweet classification is performed.

As in the case of social media accounts, not all Tweets collected with air quality-related keywords are actually related to air quality. Thus, to distinguish relevant from irrelevant Tweets we build a classifier (*Air_Quality_Tweet*) using a training set of 700 Tweets, manually labeled with respect to whether they provide information about air quality (relevant) or not (irrelevant). The same type of pre-processing is applied to the Tweets as in the case of accounts (i.e. stemming and stop-word removal) and a tf-idf bag-of-words representation is used. We also use the same classification algorithm, i.e. L2-regularized L2-loss Support Vector Machine with default parameters. Details of the classifier and its performance are shown in Table 3.

Table 3: Air_Quality_Tweet classifier details

Classifier	# examples	# relevant	# irrelevant	Precision	Recall
Air_Quality_Tweet	700	300	400	88.9%	87.2%

[21] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

After all Tweets are classified, irrelevant Tweets are removed and accounts are sorted based on the number of air quality-related Tweets they post per day. We empirically found that using a threshold of 0.4 relevant Tweets per day returns accounts that were judged by partners to be relevant for hackAIR. With this threshold, 58 accounts were returned, 47 of which were new and were not discovered with the previous methods, leading to a total number of 1435 hackAIR-relevant accounts.

4.6 Filtering inactive accounts

Since most of the hackAIR-relevant accounts are discovered based on the text of the accounts' names and descriptions, an additional step is performed to filter those of the discovered accounts that have been inactive (have not published new content) for a long period of time (we set this period to six months). To perform these filtering operations, additional API calls were required to retrieve the latest posts of each account and their timestamps. In the case of Facebook, for instance, the */userid/feed* endpoint[22] of the Graph API was queried to retrieve all public posts of the account along with their creation timestamps, while the same information is retrieved using the *Activities:list*[23] method in the case of Google+. After completing this filtering step, we are left with 493 active accounts related to air quality issues. The final list of accounts is available online[24].

4.7 Account analytics

This section presents some statistics about the discovered accounts. Table 4 shows the percentages of accounts coming from each social media platform, while Table 5 and Table 6 show the location (EU/non-EU/undefined) and language distribution (5 most popular languages) of the accounts, respectively. We notice that most accounts come from Twitter, followed by Facebook and then YouTube[25]. Regarding location, 32.6% of the accounts come from EU, while about one out of five accounts has undefined location. We also notice that English is the most common language but there are also French, Polish, German and Dutch accounts.

Table 4: Social media platforms

Social media	Percentage of accounts
Twitter	64.9%
Facebook	25.6%
YouTube	9.5%
Google Plus	0%

Table 5: Account locations

Location	Percentage of accounts
EU	32.6%
Non EU	46.5%
Undefined	20.9%

Table 6: Account languages

Language	Number of accounts
en	390
fr	47
de	18
nl	12
pl	10

[22] https://developers.facebook.com/docs/graph-api/reference/v2.10/page/feed
[23] https://developers.google.com/+/web/api/rest/latest/activities/list
[24] http://hackair-mklab.iti.gr/hackair-relevant-accounts.xlsx
[25] As explained before, Google+ returned no relevant results due to API limitations. Our search was only possible to target Google+ Users and not Google+ Pages, which would have been more appropriate.

To gain a better understanding about the geographical distribution of the discovered accounts we also plot them on a map by employing the text-based location estimation approach that was presented in D3.2 (section 2.1.2), giving as input the location field of the account. This resulted in the map of Figure 7 (zoomed in Europe in Figure 8) which is also available online[26]. We observe that many of the accounts in Europe originate from United Kingdom and in particular London. This is due to the fact that London is a highly polluted major city and that there has recently been increasing interest and engagement of the public on the topic of air pollution. In contrast, no accounts could be mapped to Norway, even though the initial collection step made use of Norwegian keywords. There are several reasons for the absence of such accounts (e.g. the classifier was too strict for Norwegian text, Norwegian air quality accounts use a different vocabulary to refer to air quality issues, Norwegian accounts do not reveal their location in the respective profile fields, etc.). To this end, further investigation will be performed in the coming period to ensure that hackAIR-relevant accounts from more geographical regions (especially from the pilots) are collected.

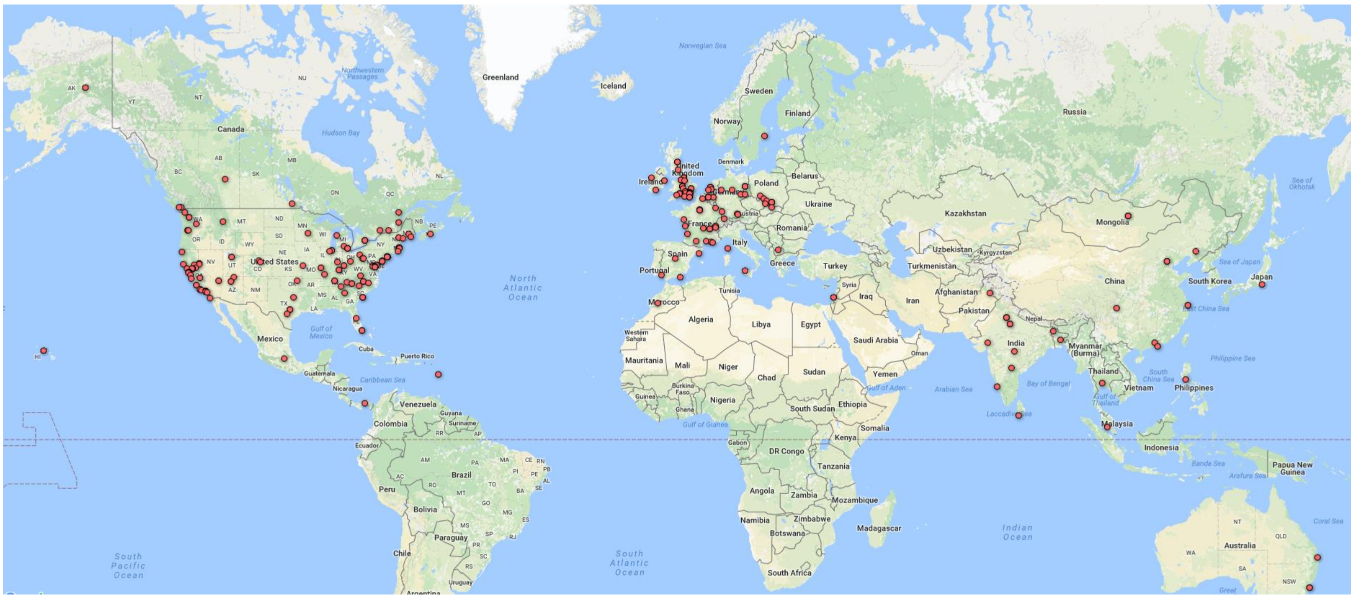

Figure 7: Estimated geographical locations of the discovered accounts

Figure 8: Estimated geographical locations of the discovered accounts (zoomed in Europe)

26 http://hackair-mklab.iti.gr/fusion/

4.7.1 Account popularity analysis

To measure the popularity of the discovered accounts we use the most established popularity metric of each social media platform, i.e. number of followers on Twitter, number of likes on Facebook, number of subscribers on YouTube. The histograms of Figure 9 show how account popularity is distributed in each social network. We observe that in the case of Twitter and Facebook there is a large number of low popularity accounts and a small number of very popular accounts, i.e. account popularity follows a power law distribution[27]. In case of YouTube, most accounts have few subscribers. The most popular accounts from each social network are shown in Figures Figure 10-Figure 13. Such accounts, in addition to the ones that we discover as part of our analysis in 5.1.3, would be very valuable to attract in the hackAIR network, since they have the potential of reaching many thousands of social media users.

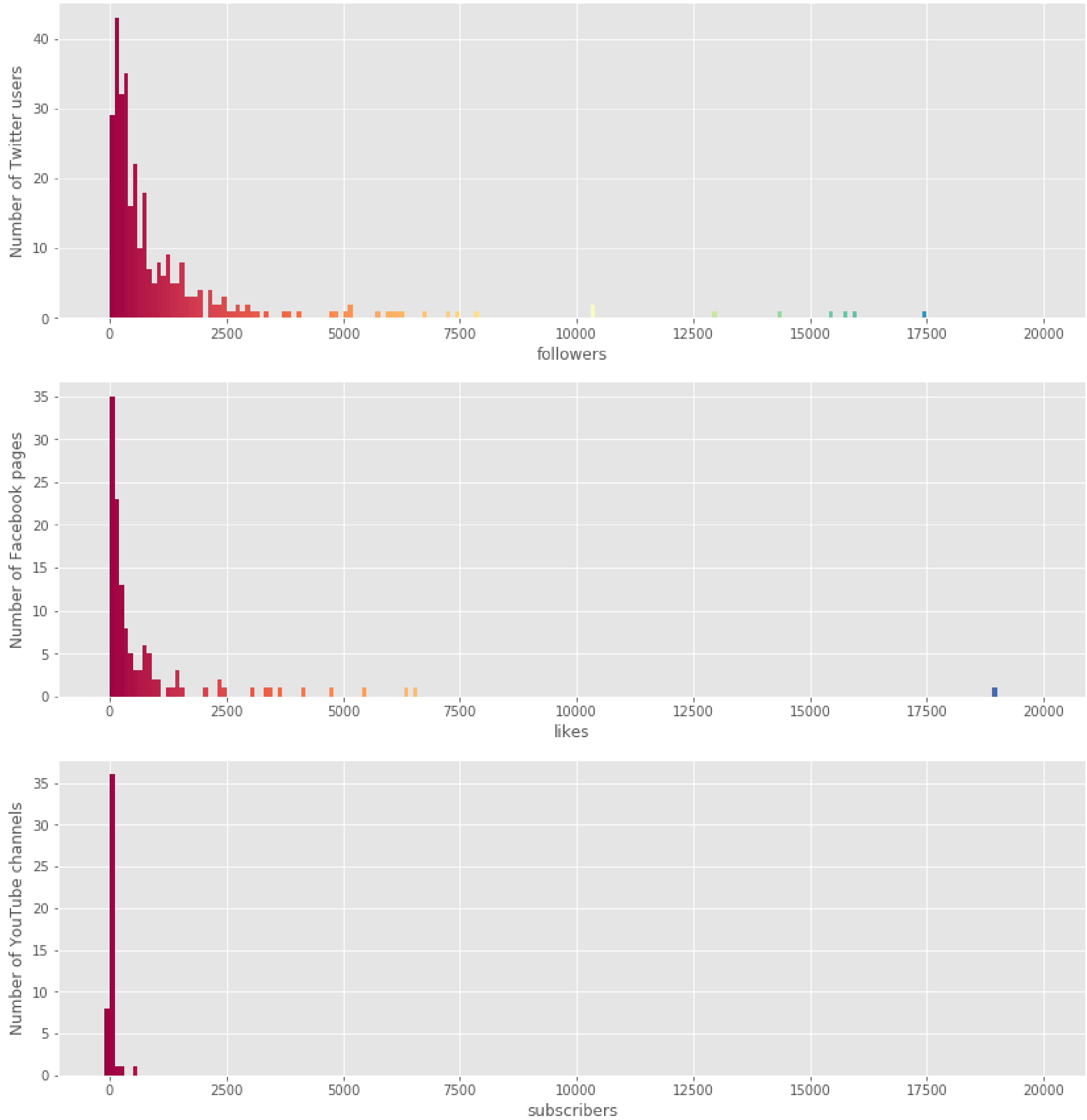

Figure 9: Account popularity histograms for Twitter (top), Facebook (middle), YouTube (bottom)

[27] https://en.wikipedia.org/wiki/Power_law

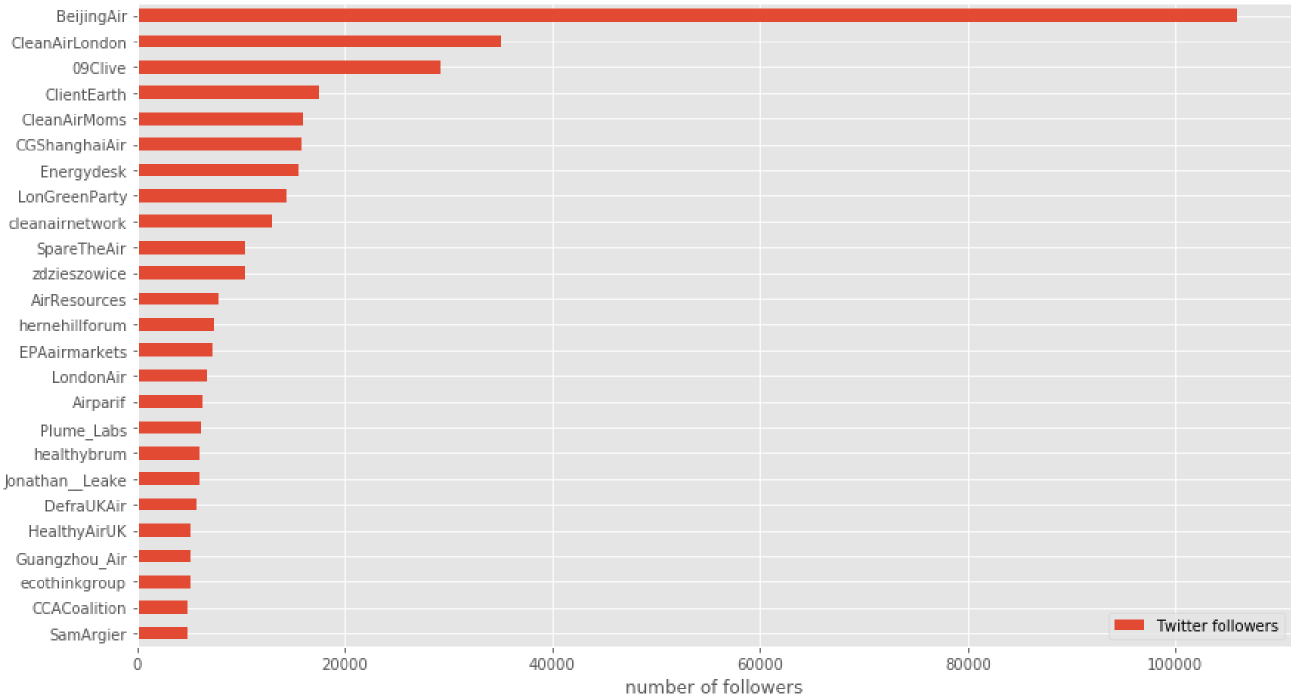

Figure 10: Top-25 Twitter accounts with respect to number of followers

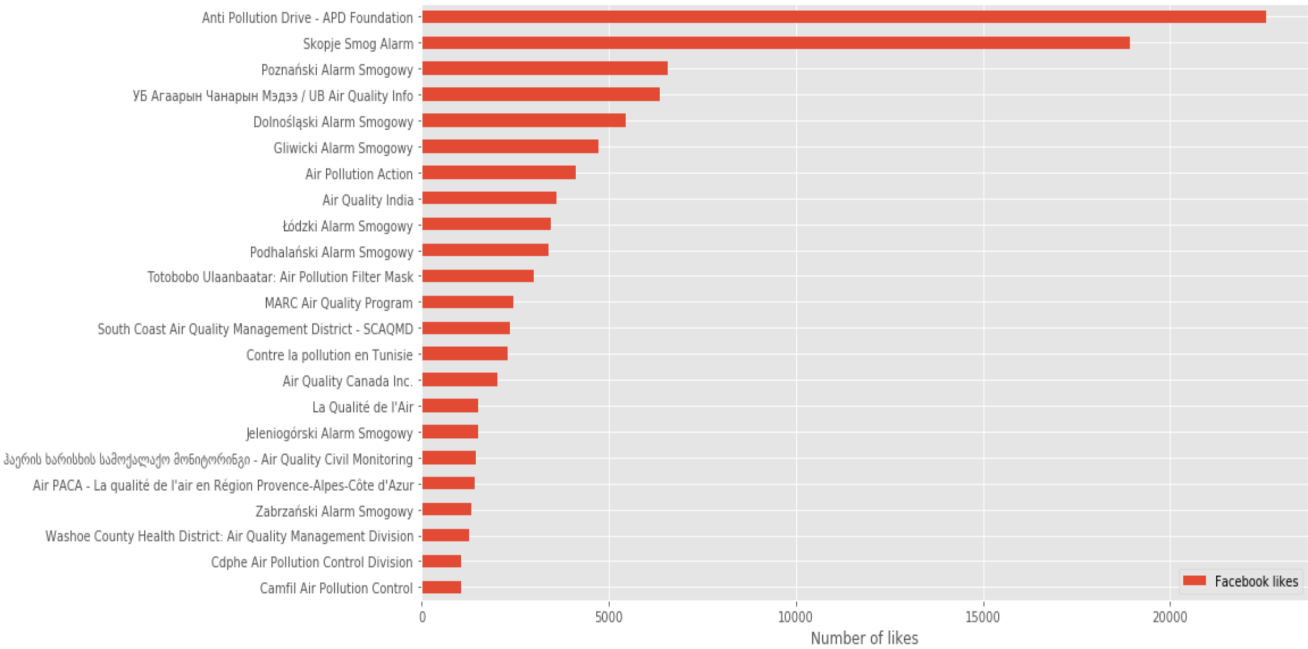

Figure 11: Top-25 Facebook accounts with respect to number of likes

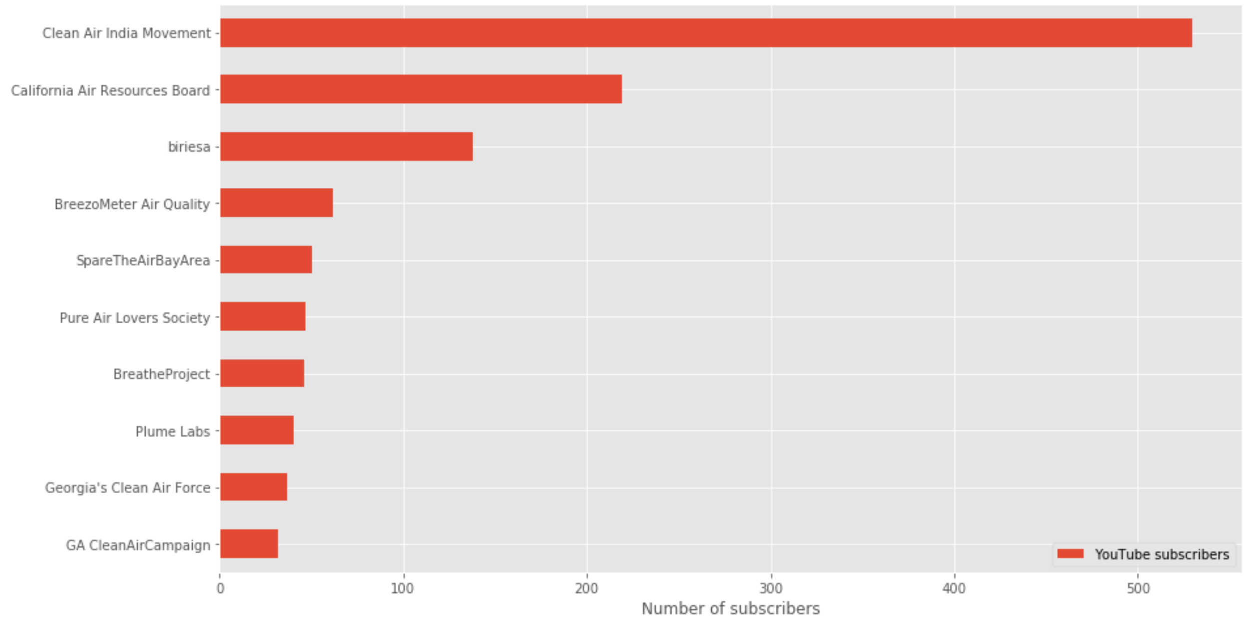

Figure 12: Top-10 YouTube accounts with respect to number of subscribers

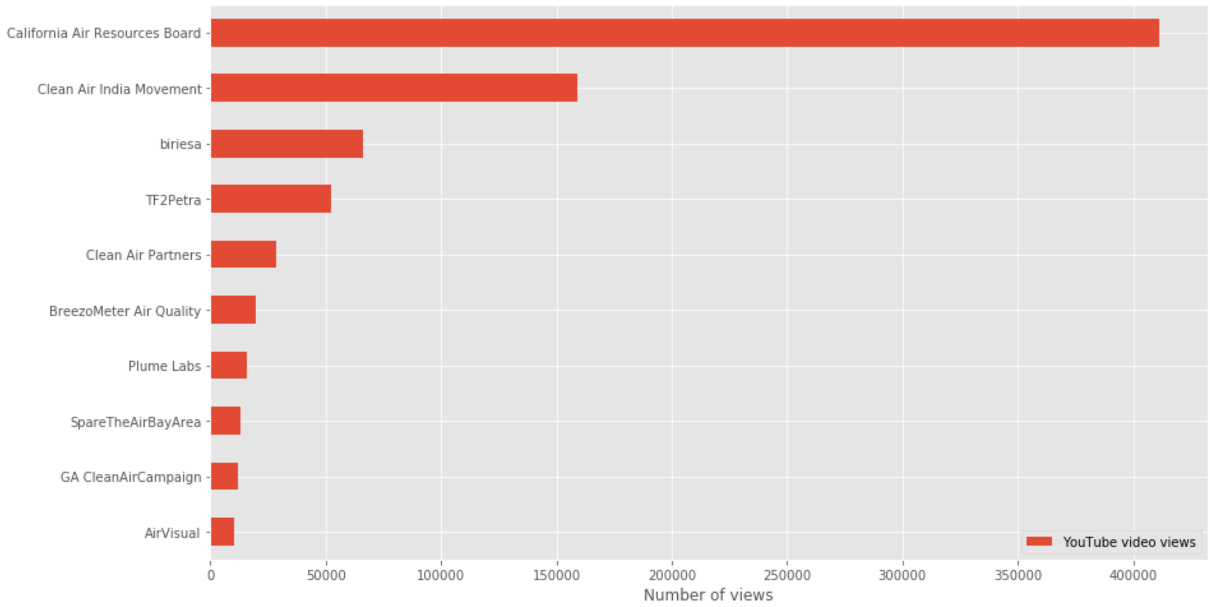

Figure 13: Top-10 YouTube accounts with respect to number of views

4.7.2 Account activity analysis

To measure the activity of the discovered accounts we use the most established activity metric of each social media platform, i.e. number of Tweets on Twitter, number of posts on Facebook, number of videos on YouTube. The most active accounts from each social network are shown in FiguresFigure 14-Figure 16. Accounts with high activity are of relevance to hackAIR, since they often share interesting news items and announcements in the areas of air quality technology and policy, and as a result the hackAIR communication managers should carefully follow their posts to discover relevant content (that could be shared through the hackAIR social media accounts).

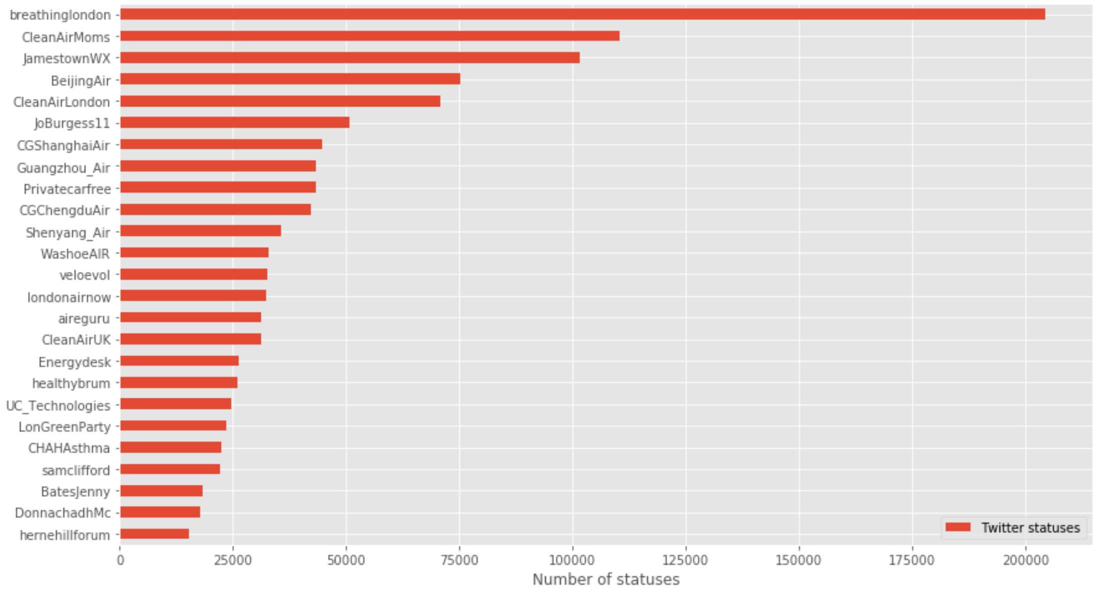

Figure 14: Top-25 most active Twitter accounts

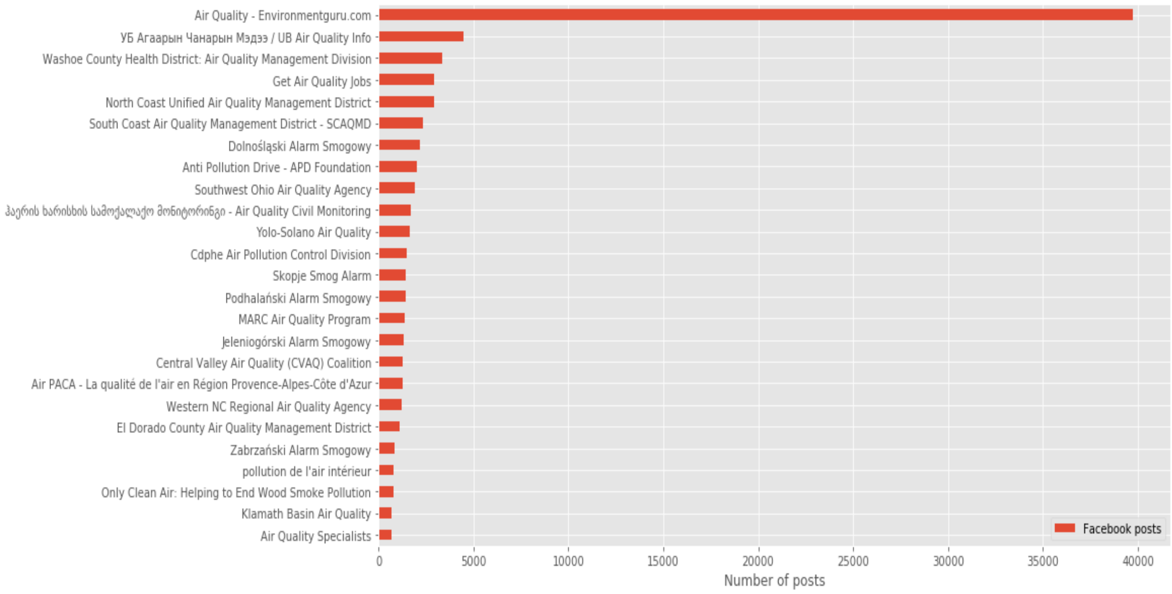

Figure 15: Top-25 most active Facebook accounts

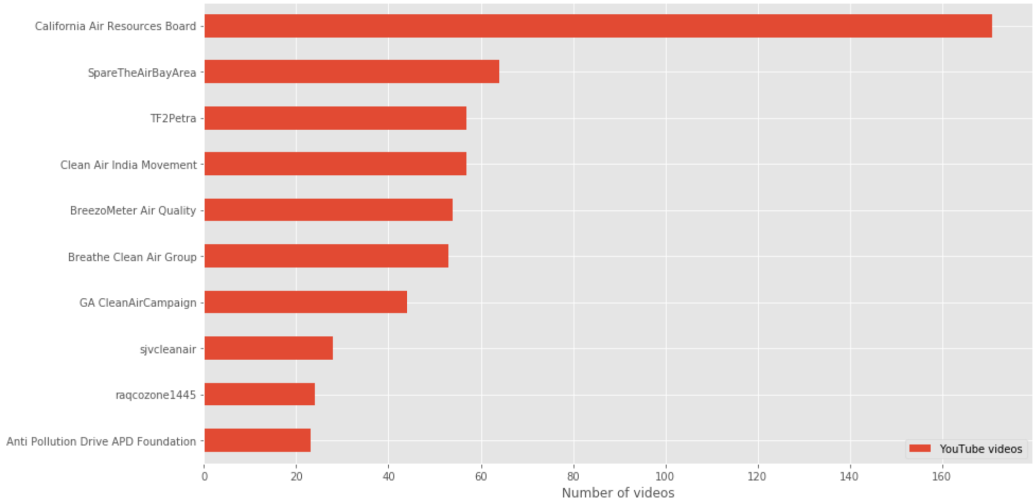

Figure 16: Top-10 most active YouTube accounts

4.8 Use cases of account discovery

In this section, we present some of the use cases of our account discovery module. Our initial goal is to find accounts from social media that are related with air quality and produce air quality related content. The account discovery module can also help detect users that could potentially act as early adopters or ambassadors of hackAIR. Such users could for example include citizens (cyclists, activists) that could be involved in generating and publishing information relevant to outdoor air pollution by submitting data through the hackAIR platform, or organizations and health associations that could organize local hackAIR workshops to build awareness, or scientific community (universities, research institutes, NGOs, independent researchers) that can use data to gain insights on air quality patterns or use hackAIR platform and communication channels for dialogue. Table 7 provides a manually created list of indicative examples of such accounts (those are a subset of the 1435 accounts that were automatically tagged by the account discovery module as relevant).

Table 7 Indicative list of accounts that could act as adopters or ambassadors for hackAIR

Name	Description
Rosalind Readhead	Environmental campaigner - Ban Private Cars in London
EllaRobertaFamilyFdn	Ella died @9 from asthma|ERFF is a Public Health Charity raising awareness of Asthma|Public Involvement member for asthma research|Campaigns for Healthy Air
Jason Pritchard	Dad, Husband, Community worker. Air pollution campaigner. Youth Mentor at @CityWestminster. Support THFC. RT does not mean agreement.
Clive Shrubsole	Senior Researcher @UCL_IEDE Indoor Air Quality Module Lead @MScHWSB, @theUKIEG, Asst Editor at IBE, Wellbeing, Climate Change Impacts. Views are my own.
Jim McQuaid	Atmospheric Scientist, sometimes airborne, branching out into atmospheric inputs for @glacier_albedo. Occasional cloud botherer @SEELeeds
Cat Scott	atmospheric scientist @ICASLeeds @CRESCENDO_H2020 & @Leeds_LEAF coordinator - interested in way natural environment & human activities interact; opinions my own
Healthy Brum	Birmingham City Council's Public Health team. Promoting healthy living and the environment across the city. #HealthyBrumhttps://t.co/1bM6VDpZqr
Lukasz Korzybski	Father, husband, environmentalist, supporter of organic farming, sustainable production, green energy and e-mobility. Freelance software engineer
Oliver Hayes	Air pollution campaigner @wwwfoecouk. Co-host @thebabblewagon podcast. Cyclist. Late night DIY enthusiast. @CambridgeUtdFC
Emilia Jane Hanna	Air Pollution campaigner for Friends of the Earth Scotland. Retweets are probably endorsements.
Julian Marshall	Professor of environmental engineering. Research: air pollution and health. Father, son, husband, brother. Bike rider.
AtmosChem Birmingham	For news of Atmospheric Chemistry and Air Pollution research from University of Birmingham

Additionally, we aim to get information from accounts by monitoring them and retrieving useful and relevant content like links or articles which are related with air quality matters and news. A list of such accounts, which was manually created based on the list of 1435 discovered accounts, is depicted in Table 8.

Table 8: Indicative list of accounts to monitor for air quality news and articles

Name	Description
ClientEarth	Justice for planet Earth. Environmental law org for forests, oceans, rights, biodiversity, climate, and air pollution in UK, EU and beyond. RTs ≠ endorsements.
CleanSpace™	Welcome to the London home of CleanSpace! Check out our local Air Maps and download the App to #SeeTheAirYouBreathe. Get the App here: https://t.co/JUArlvTxR4
Energydesk	Award-winning investigative journalism from Greenpeace. Join us on Facebook: https://t.co/sRQJImJCd8
Healthy Air Leeds	Campaigning for cleaner air in Leeds - one of the worst places for #airpollution in the UK. Tweets by Rachel (founder/co-ordinator)
Healthy Air	Cleaner air for a healthier life - Healthy Air aims to inform the public about the health threat of air pollution. Led by http://t.co/wCqPB08gGa
Plume Labs	Makers of the Plume #AirReport urban environmental forecast app. Using #OpenData & AI to beat air pollution around the world. Free on iOS, Watch & Android.
London Air	The London Air Quality Network by King's College London. Combining monitoring, modelling & toxicology to advance understanding of health impacts of pollution.
Defra Air Quality	Official, automated feed for UK Air Quality from Defra. Latest info on Pollution, Forecasts & Health Advice. For queries email aqinfo@ricardo-aea.com or visit: http://t.co/rp7cejJXki
Clean Air Network	HK's leading NGO on air pollution, the city's biggest public health crisis, killing 8 people a day.Get HK PM2.5 updates from us twice a day at @CANalert!
airTEXT	airTEXT provides air pollution forecasts for London. To signup for free air quality alerts: text airtext to 78070, or for email/voicemail visit our website.
Beth Gardiner	Environmental journalist, writing for @nytimes, @guardian + others. Currently interested in air pollution. American in London. Politics junkie. @AP alum.
Air Pollution new	Your best source of Air Pollution News on Twitter

5 Advanced audience analysis

The social media monitoring tool presented in section 3, provides a useful mechanism for tracking, over time, the impact of hackAIR's dissemination and engagement activities through social media. Given a collection that consists of the official hackAIR accounts in different social media[28] (Figure 17), the communication managers of the project can use the feed and dashboard pages (section 3.2, Figure 4 and Figure 5) to assess the audience reach and engagement with respect to hackAIR. In the feed view (Figure 18), the user can see all the posts made by the official hackAIR accounts along with number of shares/re-tweets and sort them by time of publication or popularity, while in the dashboard view (Figure 19), aggregated audience statistics such as total number of endorsements and users reached are displayed.

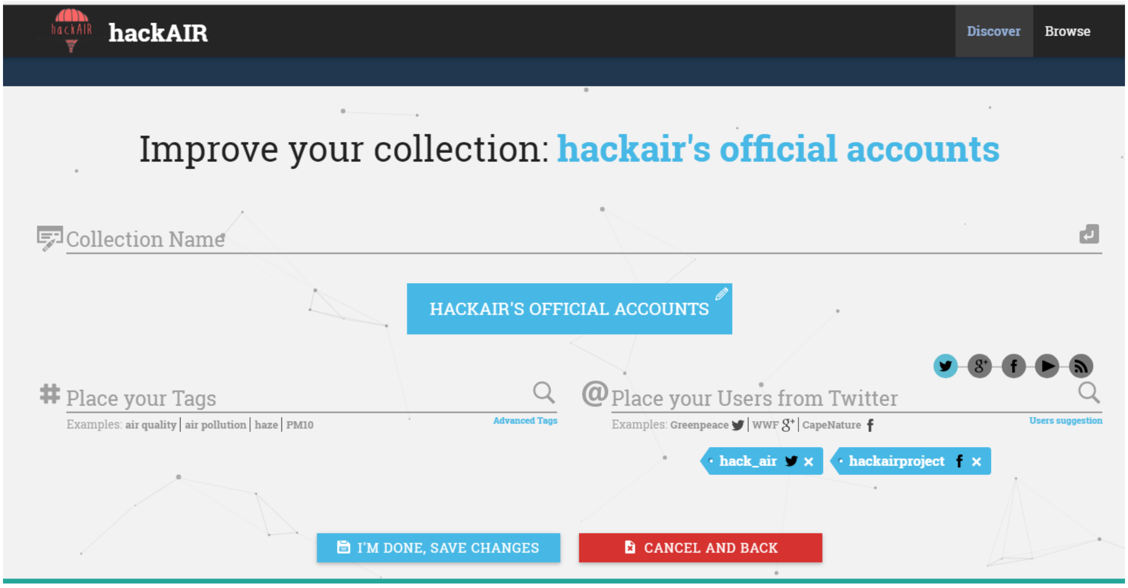

Figure 17: Definition of collection for tracking the official hackAIR account o Facebook and Twitter

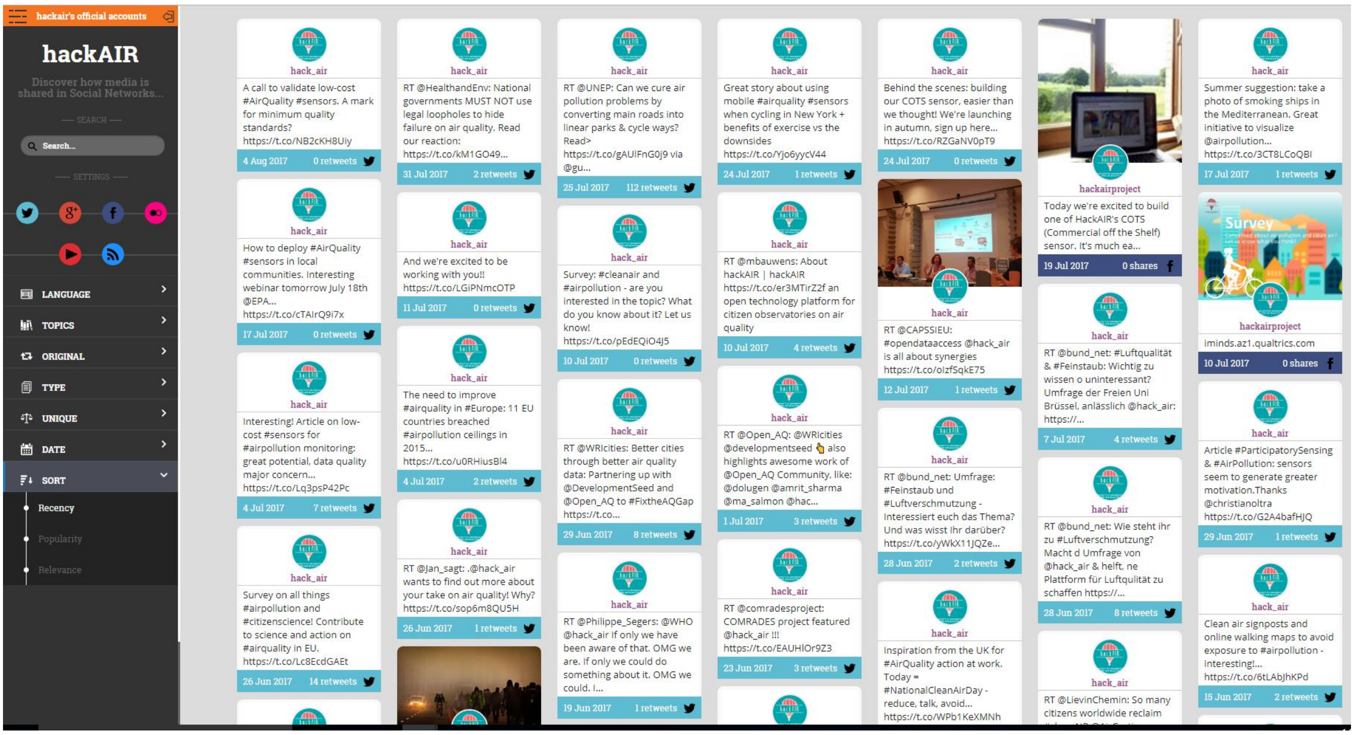

Figure 18: Feed view of the collection that tracks the official hackAIR accounts

28 Currently hackAIR is represented in Facebook and Twitter.

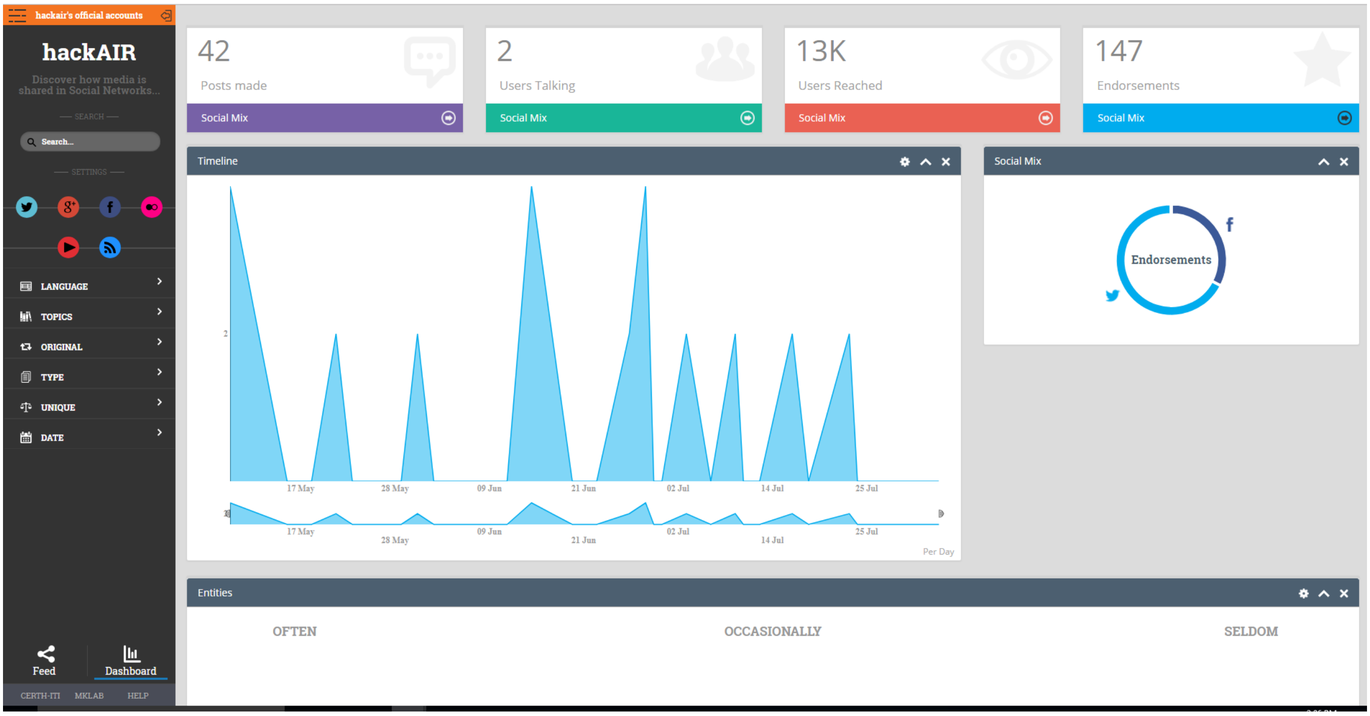

Figure 19: Dashboard view of the collection that tracks the official hackAIR accounts

Additional statistics are also provided by the native insights features offered by the social media platforms to the administrators of the accounts. Figure 20, for example, shows a snapshot from the "Reach" section of Facebook's Insights[29] page where a trend line of post reach is provided, distinguishing between organic and paid (i.e. the result of a post promotion campaign) reach, while Figure 21 shows a snapshot from the "People" section of Facebook's Insights pages where aggregated demographic data about the people who have seen content associated with hackAIR's Facebook page are provided (i.e. age/gender grouping, top countries/cities/languages).

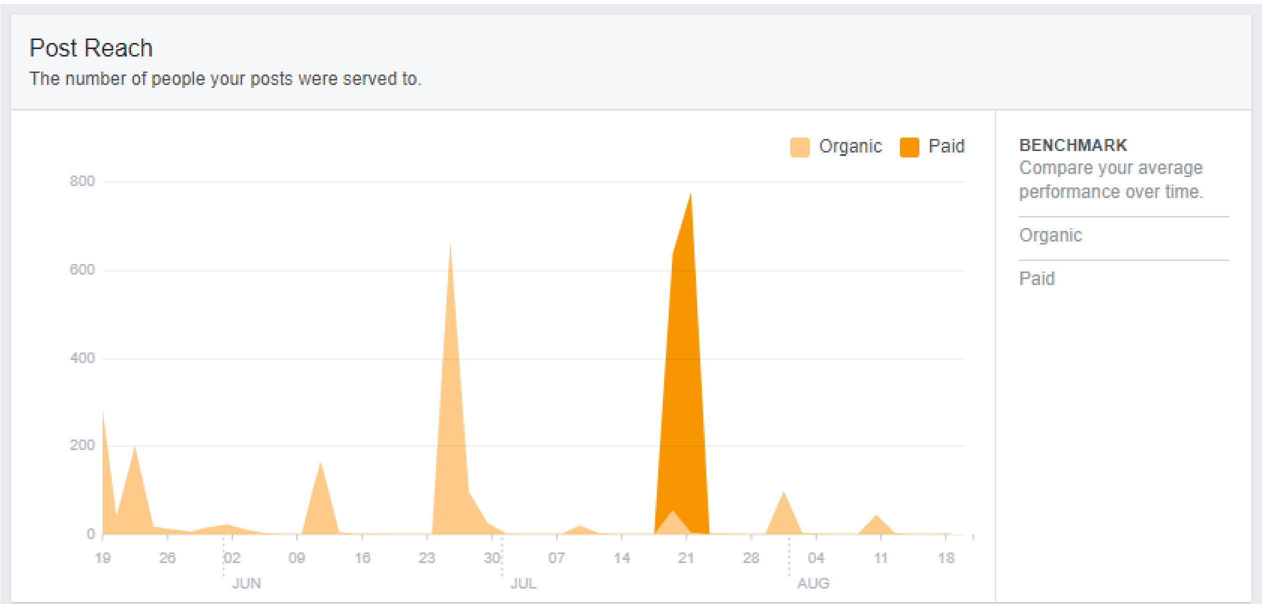

Figure 20: Post reach of hackAIR account from Facebook insights page

[29] https://www.facebook.com/help/search/?q=insights

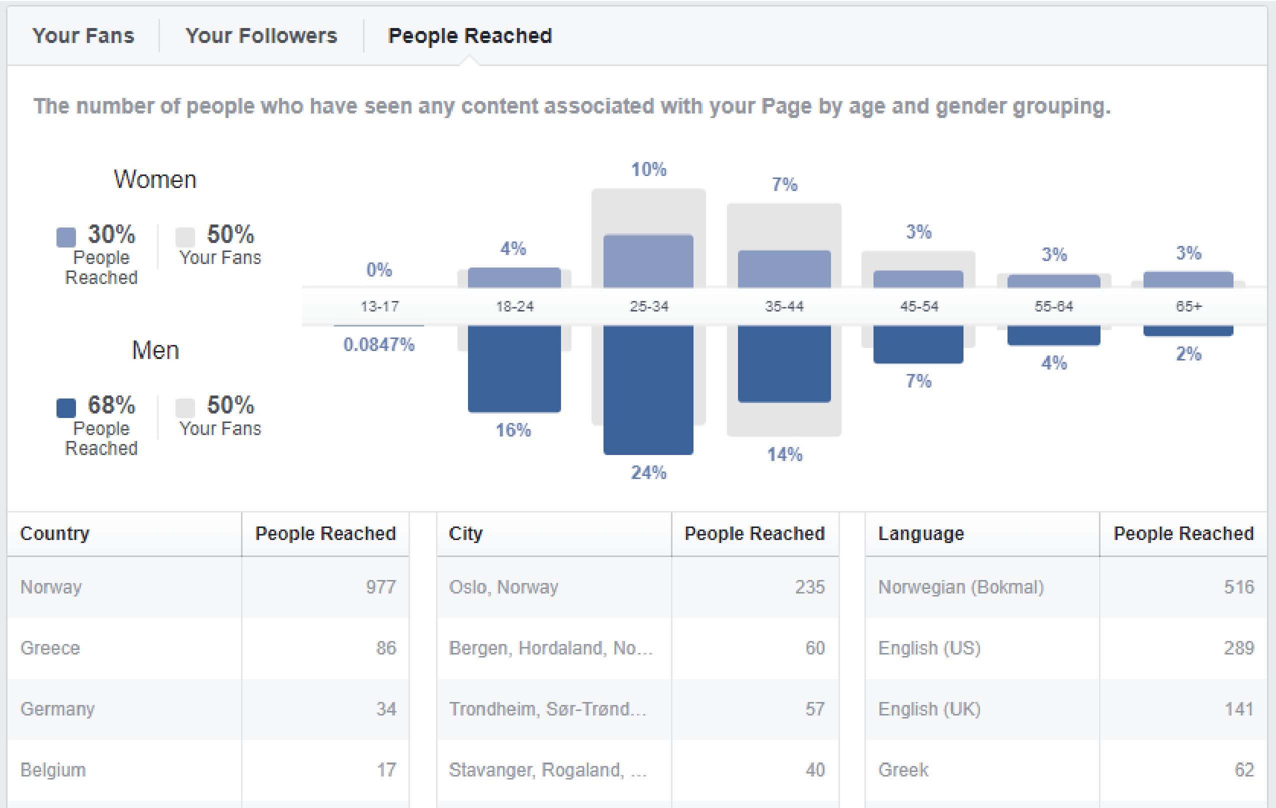

Figure 21: Audience demographics of hackAIR account from Facebook insights page

Even though the metrics provided by the current version of the hackAIR social media monitoring tool and the insights offered by the social media platforms provide a good overall idea about the state of hackAIR's popularity in social media and about the composition of its audience (in the form of aggregated demographics), they provide little or no guidance or support for implementing the hackAIR engagement strategy. In particular, they neglect valuable information that is hidden in the **structure of the network formed by the hackAIR audience**.

Following the typical structure of social networks, the network of hackAIR followers exhibits community structure (Papadopoulos, Kompatsiaris, Vakali, & Spyridonos, 2012), i.e. the nodes of the network (followers) tend to form groups that are densely connected internally (i.e. between nodes of the same group) and sparsely connected to nodes of other groups (see e.g. Figure 22). Uncovering community structure is of importance because it facilitates building an understanding regarding the dynamics of the network (e.g. information is transmitted faster within a community) and the roles that different nodes (social media accounts) have with respect to information dissemination. To this end, we devised a methodology for analyzing the structure of the network of hackAIR followers and for the discovery of communities of followers and of highly influential users in these communities. We believe that such analysis is particularly useful for the communication managers of hackAIR, offering support for designing and executing their dissemination and engagement strategy.

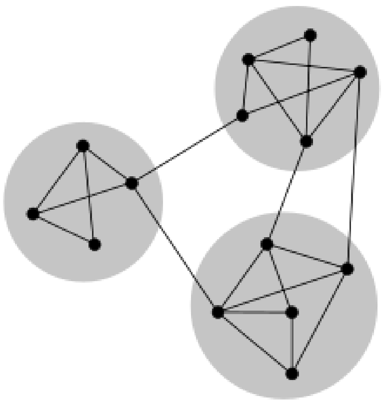

Figure 22: A network displaying community structure (source Wikipedia)

5.1 Communities in the Twitter followers network of hackAIR

In this section, we describe in detail the methodology that we applied to discover communities in the network of followers of the hackAIR official Twitter account (@hack_air). We also provide visualizations of this network where the different communities are highlighted and identify influential users in each community. This methodology is not applicable to the Facebook platform due to API limitations, i.e. we cannot obtain information about which Facebook User or Page liked a certain Facebook page.

5.1.1 Follower graph construction

A graph is a mathematical structure that is used to represent sets of objects that exhibit pairwise relations between them. More formally, a graph is an ordered pair $G = (V, E)$ comprising a set of vertices or nodes V together with a set of edges E, which are pairs of vertices (i.e., an edge associates two vertices). In our case, the vertices of the graph correspond to the followers of the hackAIR account and their followers, and the edges of the graph correspond to the follow relationships between all these Twitter accounts (i.e., the graph corresponds to the 2-hop neighborhood of the hackAIR account).

To build such a graph, the necessary data were retrieved from the Twitter API. More specifically, the *followers/ids* endpoint[30] was used, which returns a collection of user ids for every user following the specified user. The endpoint was first queried using the id of the hackAIR account (@hack_air) as input, returning the ids of the immediate followers of hackAIR. Then, to retrieve the followers of hackAIR's followers, the same endpoint was used treating each of the returned accounts as input to a new query. To give an indication about the size of the resulting network, when this procedure was applied in mid-July, it amounted to a total of 331,691 follow relationships (graph edges) between 268,123 distinct accounts (graph vertices), 374 of which are immediate followers of hackAIR[31].

Note that this data collection procedure took about 7 hours because the Twitter API allows only 60 requests per hour to the *followers/ids* endpoint. Moreover, to facilitate a more informative visualization of the graph, additional queries must be performed to retrieve the account names (e.g. @hack_air) corresponding to the retrieved account ids (e.g. 4525271542). In particular, the *users/lookup* endpoint[32] was used, which returns user objects for up to 100 users per request, as specified by comma-separated values passed to the *user_id* parameter. The Twitter API allows a significantly higher number of requests per hour to this endpoint (3600), leading to a total of 45 minutes to retrieve the names of all 268,123 accounts.

5.1.2 Community detection and visualization

To perform community detection in the graph of hackAIR followers, we used the *Louvain* algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). This is one of the most widely used algorithms and has been successfully applied to networks of many different types, including social networks sampled from Twitter (Grabowicz, Ramasco, Moro, Pujol, & Eguiluz, 2012) and LinkedIn (Haynes & Perisic, 2009). The method is a greedy optimizer of *modularity*, a measure of network structure that quantifies the quality of a community assignment by measuring how much denser the connections are within communities compared to what they would be in a random network.

The Louvain algorithm works in two steps. In the first step, the method looks for "small" communities by optimizing modularity locally. In the second step, it aggregates nodes belonging to the same community and builds a new network of which the nodes represent the first-level communities. These steps are repeated iteratively until a maximum of modularity is attained and a hierarchy of communities is produced. Although the exact computational complexity of the algorithm is not available in analytic form, it has been empirically shown to run in time $O(n \log n)$ with most of the computational effort spent on the optimization at the first level.

[30] https://dev.twitter.com/rest/reference/get/followers/ids
[31] A reduced version of this graph (after filtering nodes with only a single connection), which was used for the rest of the analysis, can be found here: http://hackair-mklab.iti.gr/simplified_network.zip
[32] https://dev.twitter.com/rest/reference/get/users/lookup

Before applying the Louvain algorithm on the hackAIR follower graph, we first removed leaf-nodes (i.e. Twitter accounts that are connected to only a single account) to simplify the graph and to generate an easier to read visual representation. This led to a smaller (and denser) graph of 32,051 nodes and 95,582 edges. When the Louvain algorithm was applied on this graph, 13 different communities were discovered (0.545 modularity score). We used the implementation offered by the Gephi[33] open-source graph visualization software with default parameters.

Figure 23 depicts a visualization of the graph (obtained using Gephi) where the layout was determined by the ForceAtlas2 (Jacomy, Venturini, Heymann, & Bastian, 2014) graph layout algorithm, nodes are colored according to community assignment and their size is proportional to their edge degree (i.e. total number of incoming edges). By inspecting the most influential nodes (those with higher degree) of each community, the following 10 relevant communities of Twitter accounts were identified (labeled in Figure 23): **1) science-innovation, 2) smart-green cities, 3) technology, open data and sensors A, 4) technology, open data and sensors B, 5) UK and general air pollution accounts, 6) German accounts, 7) Belgian and Norwegian accounts, 8) French accounts, 9) Irish accounts, 10) Greek accounts**. The hackAIR account is located at the center of the graph and was assigned by the algorithm to the community of UK and general air pollution accounts. We have also created a very similar web-based view of this graph to facilitate the interactive exploration of the hackAIR two-hop neighborhood. This web-view is available online[34].

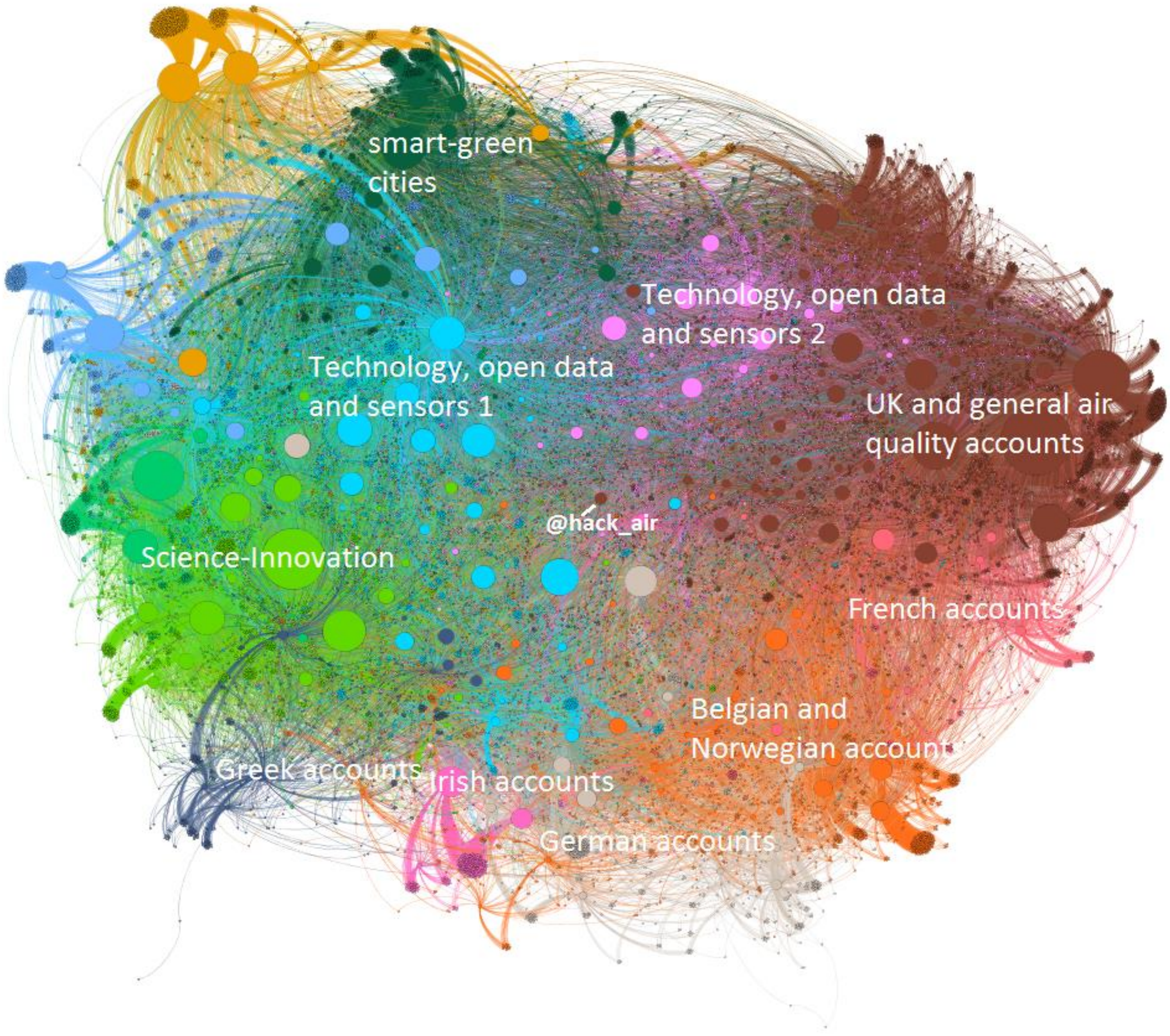

Figure 23: Visualization of different communities in the hackAIR followers network

[33] https://gephi.org/

[34] http://hackair-mklab.iti.gr/hackair-network/: At the time of writing this deliverable, the implementation of the visualization is based on the *sigma.js* graph visualization library, which limits the number of nodes and edges that can be shown in interactive mode, and as a result a reduced version of the hackAIR network is shown. In the coming period, we plan to switch into a WebGL-based library that will make possible the interactive visualization and exploration of graphs with tens of thousands of nodes.

Figure 24 zooms into the science-innovation community, allowing a more detailed inspection of the community's structure such as the most influential users and connections between them. Similar zoomed-in views have been generated for all the discovered communities and are provided in the Appendix (Figures Figure 26-Figure 34).

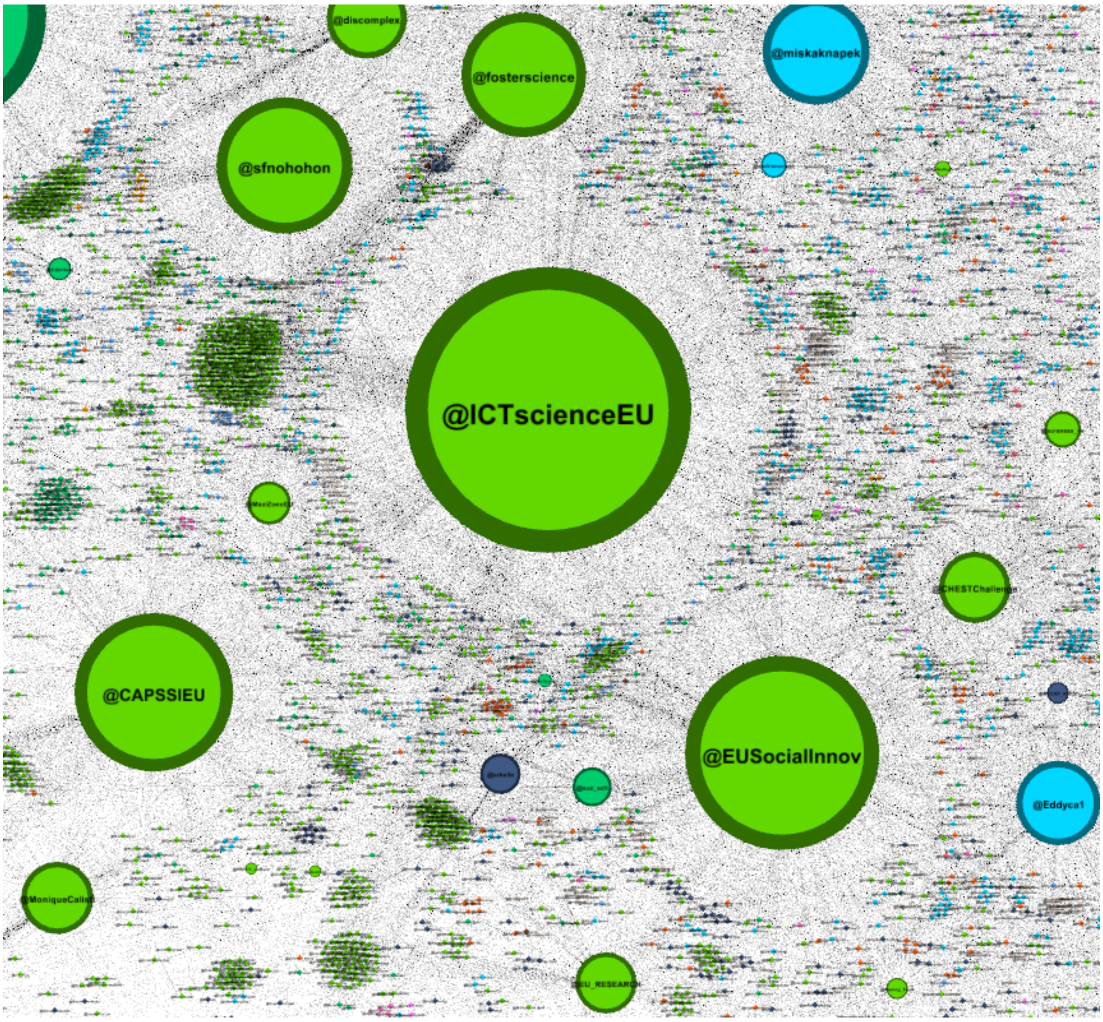

Figure 24: The science-innovation community of the hackAIR followers network

5.1.3 Influential Twitter accounts

In this subsection we try to detect influential nodes among the nodes of the hackAIR follower graph.

5.1.3.1 Influential accounts in the hackAIR follower graph

Using information from the graph, we could locate the most influential accounts, i.e. accounts that are followed by many others (and therefore likely to exert influence on them). To this end, we first filtered nodes based on their *in-degree*, i.e. the number of incoming edges, keeping only nodes that were followed by at least 100 accounts. We then manually inspected the remaining nodes (216) in order to select those that correspond to individuals. To limit the amount of manual effort, we used the reduced graph described in section 5.1.2 (i.e. the graph after removing all leaf-nodes). The resulting list contains 55 accounts. Among these accounts, there are researchers, journalists, innovators, entrepreneurs and generally users that post about issues related to the environment, air quality, green infrastructure and sustainability. These influential users are shown in Table 13 in the Appendix. Notice that due to the way that the input graph was constructed (as described in 5.1.1), all discovered influential users are followers of the hackAIR account (as they have higher number of incoming edges compared to their followers, for which no followers were collected). Although discovering influential users out of the existing follower base of hackAIR is useful for tailoring the social media strategy of the project, we were further interested in relevant influential accounts that do not follow the hackAIR account (and hence could be the targets of communication efforts with the hope of growing the project follower base). This effort is described in the next paragraph.

5.1.3.2 Influential accounts that do not follow hackAIR

To find influential users that appear in the graph but do not follow the hackAIR account we need information about the number of followers of these users. This information was already retrieved when we queried the *user/lookup* endpoint in section 5.1.1. Due to the fact that a large amount of these users may be irrelevant with air pollution topics, we make the following assumption regarding relevance: the relevance of an account for hackAIR is highly dependent on the number of hackAIR followers that this account follows (since these accounts are highly likely to regularly post about air quality issues and news).

With this in mind, we first filtered accounts with less than 1000 followers (in that case 1000 was used instead of 100, since the initial set of accounts was much larger and contained much more influential accounts), and then also filtered accounts that followed less than 10 hackAIR followers. The resulting list comprised 270 accounts, which we manually inspected in order to identify the ones that correspond to individuals. Table 14 in the Appendix presents the resulting 44 accounts. This list can serve as a valuable resource for the hackAIR communication managers, who can directly approach these accounts with the goal of engaging them in the hackAIR outcomes. This will happen regularly in the coming period, and to further assist in this process, this list of influential users will be regularly updated by the social media monitoring tool.

5.2 Reach of hackAIR account

To get more insights about the number of users that hackAIR posts can reach, we calculate the maximum reach of the hackAIR account (how many users can see a tweet if all followers retweet it) and we try to detect the most active followers and their corresponding communities. The hackAIR Twitter account has 374 followers. From the previous section, it was found that 268,123 accounts follow these 374 users. Therefore, if every hackAIR follower retweeted its tweets, these could potentially reach 268,123 accounts. In practice, however, only a small number of followers retweet posts (this is a general observation and not particular for hackAIR). To quantify the actual retweeting activity around hackAIR, we query the statuses/user_timeline[35] endpoint which returns a collection of the most recent Tweets posted by a certain user (hackAIR in our case). At the end of July 2017, the hackAIR account timeline had 305 tweets, 136 of which were retweeted tweets from other users and 169 were original tweets. Furthermore, we queried the *statuses/retweets/:id*[36] endpoint, which returns a collection of the 100 most recent retweets of the Tweet specified by the id parameter, and used the ids of the above 169 tweets. Results showed that 61% of hackAIR tweets are retweeted by at least one account. Also, 124 of all followers have retweeted at least one tweet of the hackAIR account. Table 9 presents the top 15 accounts in terms of number of retweeting the tweets of the hackAIR account. Regularly monitoring this list can help communication managers identify the type of content that is mostly appealing to certain followers so that they can tune their future posting activity.

Table 9: Top 15 retweeters of hackAIR account

Username	Retweets	Community
@AFellermann	22	German accounts
@CAPSSIEU	15	Science – Innovation
@CleanAirLondon	13	UK and general air quality accounts
@CleanAirUK	11	UK and general air quality accounts
@wiebkehere	8	Technology, open data and sensors 1
@codeforS	6	German accounts
@CERTHellas	4	Greek accounts

[35] https://dev.twitter.com/rest/reference/get/statuses/user_timeline
[36] https://dev.twitter.com/rest/reference/get/statuses/retweets/id

@greenfabbxl	4	Belgian-Norwegian accounts
@ClairCity	4	UK and general air quality accounts
@Philippe_Segers	4	Science – Innovation
@airqualityindia	4	UK and general air quality accounts
@cliffordious	4	UK and general air quality accounts
@PSyropoulou	3	Greek accounts
@UTconnect	3	Technology, open data and sensors 1
@Open_AQ	3	UK and general air quality accounts

In a further analysis, to find the most actively retweeting communities, we calculate for each community the number of accounts that retweet hackAIR tweets and the number of retweets. Results for the most active communities are presented in Table 10. According to it, communities comprising general air quality accounts, science and innovation accounts, and German accounts are the three most active in terms of spreading the hackAIR posts. Knowing which communities engage most with the posts of hackAIR could help communication managers in two ways: a) reinforce already active communities by increasing the number of posts about topics they have already retweeted, b) engage more actively with the remaining communities by diversifying the topics posted (e.g. post more topics about smart cities and air quality to engage the *Smart-green cities* community which currently ranks eighth). An additional observation that one may make by seeing the data is that the language of posts may play a role in the spread of the message. For instance, Belgian-Norwegian, Greek and French communities rank below English and German ones. This may indicate that a multi-lingual communication strategy should be considered for maximizing the geographical spread of hackAIR posts across Europe.

Table 10: Most active communities based on the number of retweeters

Community	Accounts retweeted	Retweets
UK and general air quality accounts	32	69
Science – Innovation	20	44
German accounts	15	45
Belgian-Norwegian accounts	14	19
Technology, open data and sensors 1	13	27
Smart-green cities	11	16
Greek accounts	5	12
French accounts	4	6

The final part of our reach analysis is based on the analysis of the distribution of retweets for hackAIR posts and their relation with the actual reach they achieve on the hackAIR follower network. The left part of Figure 25 illustrates the distribution of retweets for hackAIR posts. It appears that the majority of hackAIR posts receive zero or one retweet, while there are only few posts that were retweeted heavily (>10 times). However, analysis of the number of retweets on its own does not paint the whole picture. For that reason, we made use of the hackAIR follower graph in order to compute the maximum reach for each post. This was done by counting the total number of nodes on the network that could potentially read the post. This is equal to the union of the neighborhoods (sets of neighboring nodes) of the hackAIR account and of all accounts that retweeted the post. Compared to the number of retweets, this is a much more accurate measure of the audience size for hackAIR, since each time an account retweets a post, all its followers may potentially see it[37]. The right part of Figure 25 provides a much more insightful view into the reach of hackAIR posts. This is a scatter plot of the reach for each post (each post corresponds to a dot) in relation to the number of retweets that it achieved (*x* axis). Based on this plot, one may categorize hackAIR posts in three groups:

- *Normal:* These are posts that achieve the "expected" reach given the number of retweets. They are the blue dots appearing at the bottom left part of the plot and exhibit a roughly linear relation between reach and number of retweets.
- *High performers:* These are posts that exhibit very high reach even though their numbers of retweets are in the same order of magnitude as the ones achieved by the Normal posts. They are the green dots appearing at the top part of the plot.
- *Low-reach:* These are posts that only achieve low or moderate reach even though their numbers of retweets are higher than the ones of the Normal or even the High Performers.

Further insights are gained by inspecting some examples of posts from each of the above categories. Table 11 lists four examples from each category along with the number of retweets and reach that each of them achieves. A first noteworthy observation is that for the same number of retweets, High performers can reach tens of times more accounts compared to Low-reach posts (e.g., compare the second High performer that achieved a reach of over 40K accounts with just 7 retweets with the fourth low-reach post that achieved a reach of only 553 for the same number of retweets. By inspecting more closely the High Performers, we observed that in all cases, their very high reach was due to the fact that they were picked up and retweeted by the highly influential @CleanAirLondon account, which has more than 37K followers. Three out of the four example such posts contain links to relevant articles, while the fourth one informed about two of the hackAIR co-creation workshops. A key learning from this analysis is that it will be very beneficial for hackAIR to attract the interest of highly influential accounts such as @CleanAirLondon and manage to get them to spread the news about the project. Thanks to the analysis of section 5.1.3 and the resulting lists of influential accounts (Tables Table 13 and Table 14), it is now possible for the hackAIR communication managers to focus their promotion efforts on such high-profile accounts.

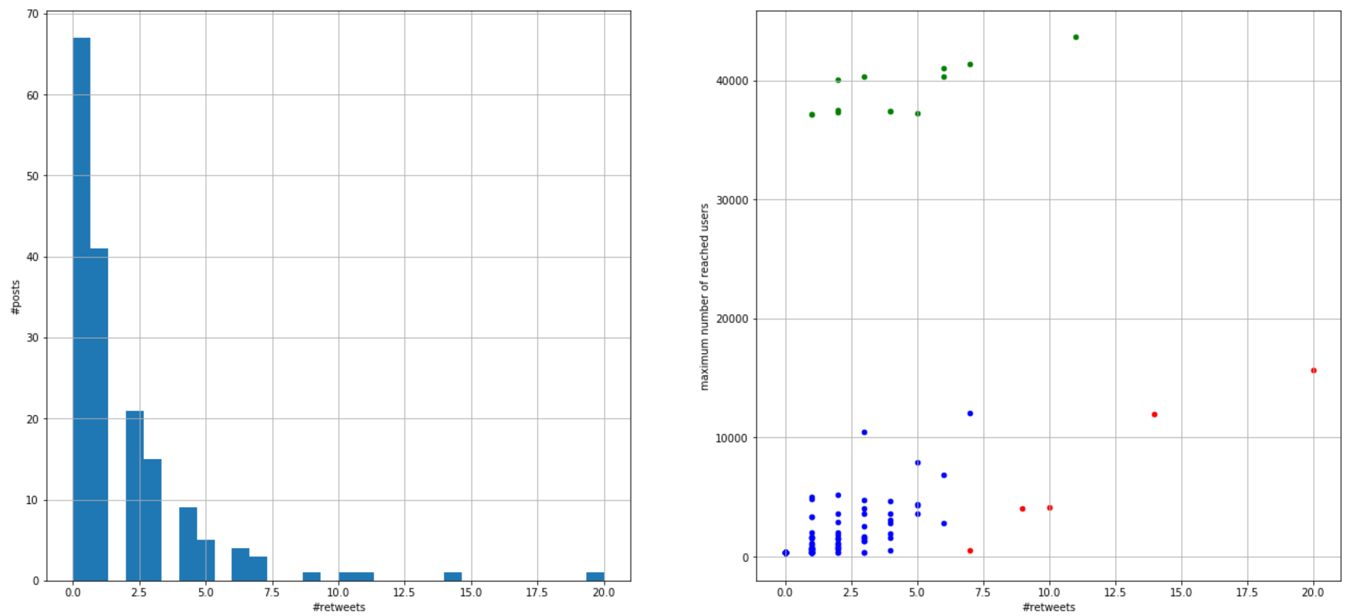

Figure 25: Retweet and reach analysis of 305 hackAIR posts. Left) Histogram of posts' retweets; Right) Scatter plot of posts' reach on the hackAIR follower graph versus their retweets.

[37] In reality, only a fraction of follower accounts really sees a given post that is tweeted by the followed account. The actual number of times that a tweet is seen by other accounts is more typically referred to as *number of impressions* and is reported by the native Twitter analytics tool (only for the last 28 days). In the case of the hackAIR account, the number of impressions for the last 28 days (3-30 August 2017) amounted to 6,113.

Table 11: Indicative examples of posts from each category (Normal, High performers, Low-reach).

Tweet	Retweets	Reach	Category
Interesting! Article on low-cost #sensors for #airpollution monitoring: great potential, data quality major concern... https://t.co/Lq3psP42Pc	7	12,061	Normal
We're busy! Sensor testing, platform development, user engagement,... Action on #airquality in GR, NO, GE, BE, NL https://t.co/ZqHWKnvpuf	2	1,055	Normal
Sept 12: Workshop Internet for #EnvironmentalMonitoring w/ @NILU_now @CERTHellas @DraxisEnv https://t.co/IpQ5YGYqND https://t.co/nn77opdeZ8	1	877	Normal
How #airpollution affects your health - infographic on the main pollutants and their health impacts https://t.co/rrzvpb4Tdt	0	374	Normal
Great story about using mobile #airquality #sensors when cycling in New York + benefits of exercise vs the downsides https://t.co/Yjo6yycV44	11	43,663	High performer
Benefits of #opendata. No clear figures on #airquality: any research on how open data helps fighting #airpollution? https://t.co/f4gVachmhM	7	41,352	High performer
Health risks of #airpollution: Living near a freeway is equivalent to 10 daily passively smoked cigarettes.@psmonks https://t.co/UNLwuduQJ8	6	40,977	High performer
And you can discuss #airquality #monitoring with @AFellermann in 2 co-creation workshops, 20 Oct & 10 Nov in Berlin https://t.co/9Sr06FbuiL	3	40,301	High performer
Survey on all things #airpollution and #citizenscience! Contribute to science and action on #airquality in EU. https://t.co/Lc8EcdGAEt	14	11,973	Low-reach
How to deploy #AirQuality #sensors in local communities. Interesting webinar tomorrow July 18th @EPA... https://t.co/cTAIrQ9i7x	10	4,172	Low-reach
How #OpenData platforms, low-cost sensors & mobile apps are powerful tools for better #AirQuality... https://t.co/BCqPdBJDiI	9	4,031	Low-reach
hackAIR project presented by Paulien Coppens & @VeeckmanCarina at #UnpluggingData for #smartcity regions workshop... https://t.co/hNh34LAiQD	7	553	Low-reach

6 Conclusions and future work

The deliverable presented the outcomes of our work on the development of social media monitoring tools with the goal of supporting and assessing the engagement strategy of the project and its communication and dissemination activities. After presenting the core social media monitoring technology that was used as basis for our work, the deliverable focused on two new capabilities that were considered to be of importance for the hackAIR engagement and communication strategy:

- Discovery of accounts that are of relevance to the project based on the text description of their profile and the text of their posts.
- Analysis of hackAIR audience with the help of a community detection algorithm and the use of graph analysis to identify influential accounts within and outside the follower base of the hackAIR Twitter account.

At the time of delivering this report, the social media monitoring tool of Section 3 is set up and monitors social media activity around topics of relevance for hackAIR. In addition, the newly developed capabilities described in Sections 4 and 5 have been applied on actual data of relevance to hackAIR, and their results were made available to hackAIR partners and their further use in the next period of the project has been discussed.

In particular, following up on a discussion with the project communication manager (ONSUBJECT) and engagement strategy experts (VUB), the following actions are planned[38].

Identification of hackAIR ambassadors: Through the social media monitoring tools influential users can be identified, who are already part of the hackAIR network or who are not yet. In respect to the engagement strategy in WP6, pilot coordinators and local communication managers, in collaboration with CERTH, will rerun the processes described in Section 5 to identify ambassadors on a regular basis (once per quarter).

Engagement via following relevant accounts. Once the hackAIR platform is launched in autumn 2017, the hackAIR social media accounts will follow all accounts on the list of relevant social media accounts that was produced by the method of Section 4. This will result in a notification to these users, and many of them are expected to take the opportunity to check out hackAIR's profile and potentially follow it back. The lists of automatically discovered relevant accounts will be regularly updated in order to reinforce this type of engagement.

Engagement via direct communication. The hackAIR communication manager will select and contact a subset of the most relevant accounts directly (either by direct message or @mention) and invite them to try out the hackAIR platform. The selection of such accounts will be based on how likely it is for them to contribute to the pilots.

Assessment and communication feedback. Audience analysis activities similar to the ones presented in Section 5 will be periodically carried out (once per quarter) and their results reviewed during the regularly scheduled engagement calls: in that way, it will be possible to focus on actionable insights derived from such analysis and use these insights to improve the communications in the pilots and the overall project.

Discovery of hackAIR-newsworthy articles from social media. This will provide hackAIR social media account managers with content that is worth sharing and that could eventually lead to an increased follower base. In addition, it will help the consortium stay up-to-date with the latest developments and trends in the area of air quality research, technology, activism and policy. From a technical point of view, newsworthy content discovery will be based on the monitoring of shared links to articles from selected lists of accounts and on devising relevance models that use the content of the articles and their popularity (e.g. number of likes, retweets) to rank them before presenting them to the hackAIR social media managers.

[38] Given that WP6 (T6.3) activities are concluded with the delivery of this report and the accompanying tools, the follow-up activities will be carried out within ongoing tasks T5.3 (for technical improvements that will be required as a result of partners' feedback) and T7.3-T7.4 (for interactions and analysis related to pilot-specific data and results).

Improvements on advanced account discovery. Future technical work will focus on improving the quality, relevance and comprehensiveness of automatically produced results. Initially, we aim to improve the classifier of relevant-irrelevant accounts by taking into consideration the text content of their posts (in addition to the text description in their profile). This would lead to even more accurate results and could also enable us to discover more keywords related to air quality, which could in turn be used to fetch more candidate accounts.

Another improvement concerns the extension of the user discovery process based on tweets (as presented in section 4.5) to leverage tweets from all over Europe and the world. This will make possible to collect air pollution-related content from more sources around Europe and the world. Finally, we plan to further use a multimedia geotagging[39] module to make estimations about the location of accounts by analyzing the text of their posts or other metadata, e.g. the description or the name in their profile.

[39] https://github.com/MKLab-ITI/multimedia-geotagging

7 Appendix

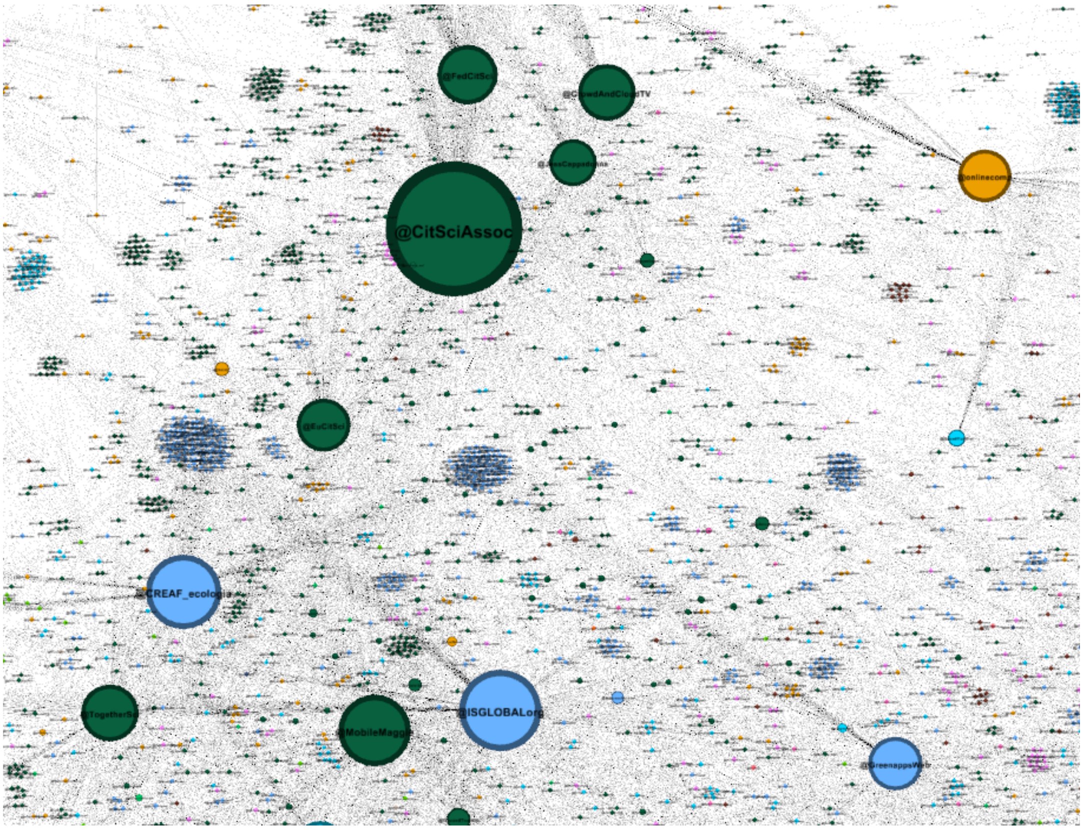

Figure 26: The smart-green cities community of the hackAIR followers network

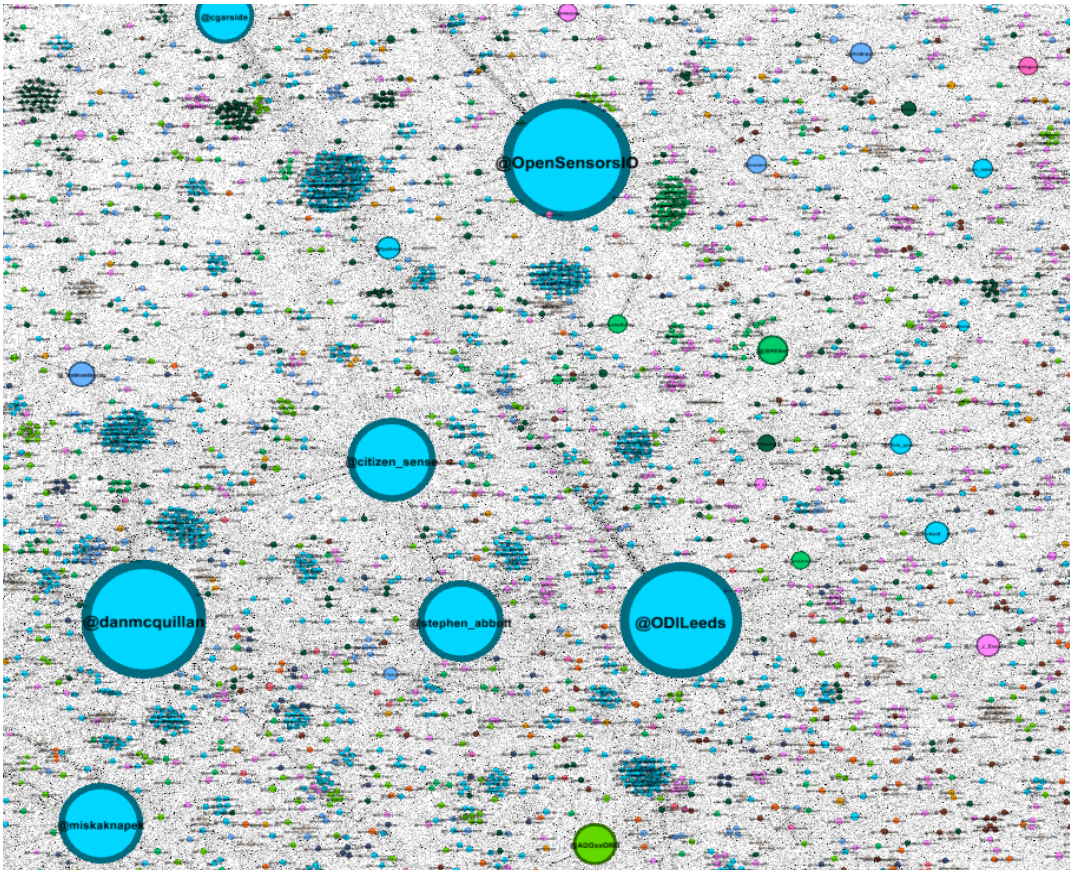

Figure 27: technology, open data and sensors 1 community of the hackAIR followers network

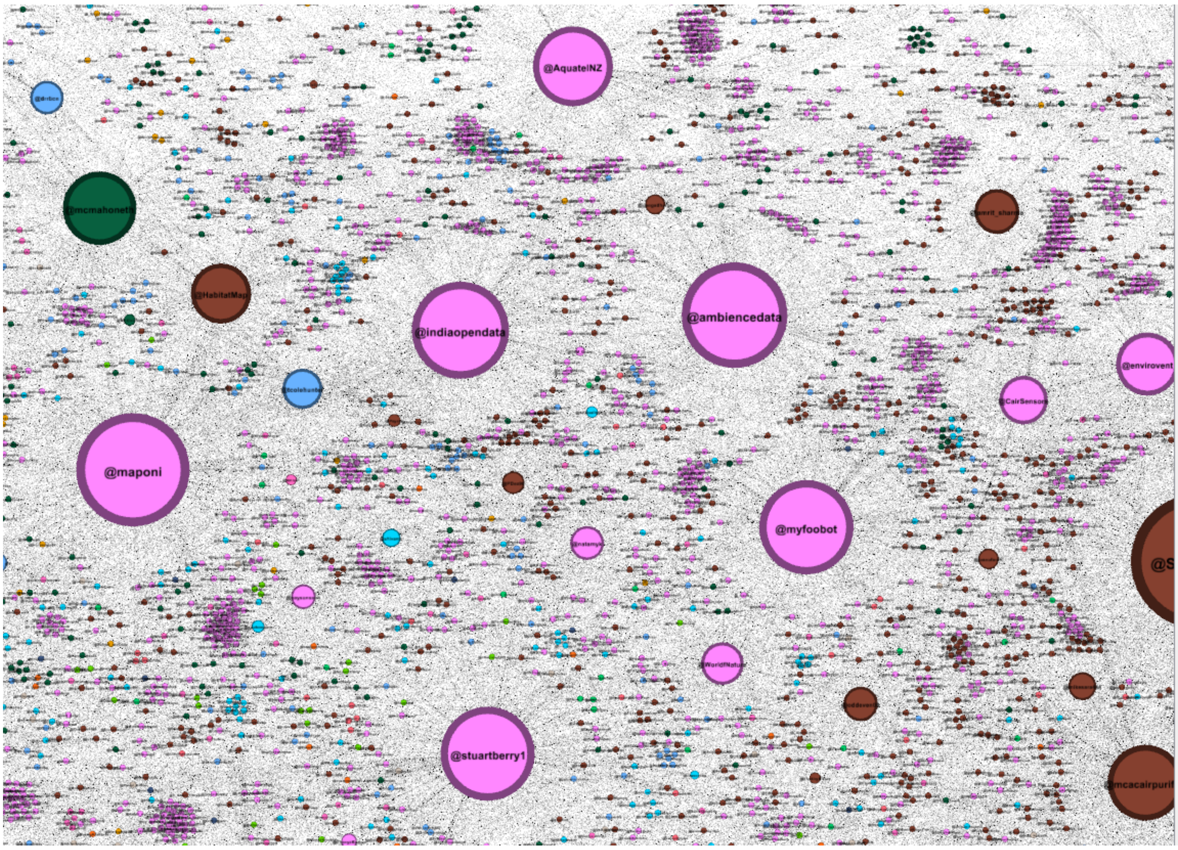

Figure 28: technology, open data and sensors 2 community of the hackAIR followers network

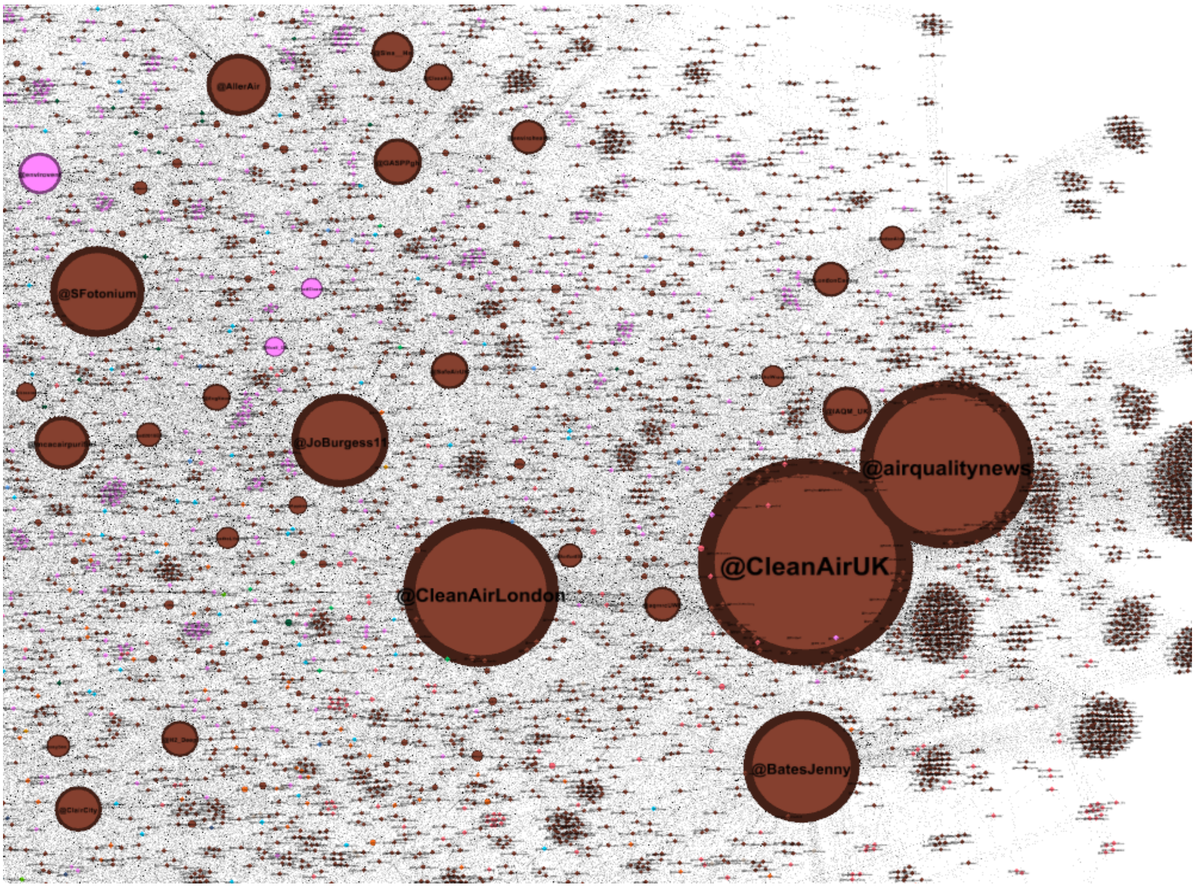

Figure 29: UK and general air quality accounts community of the hackAIR followers network

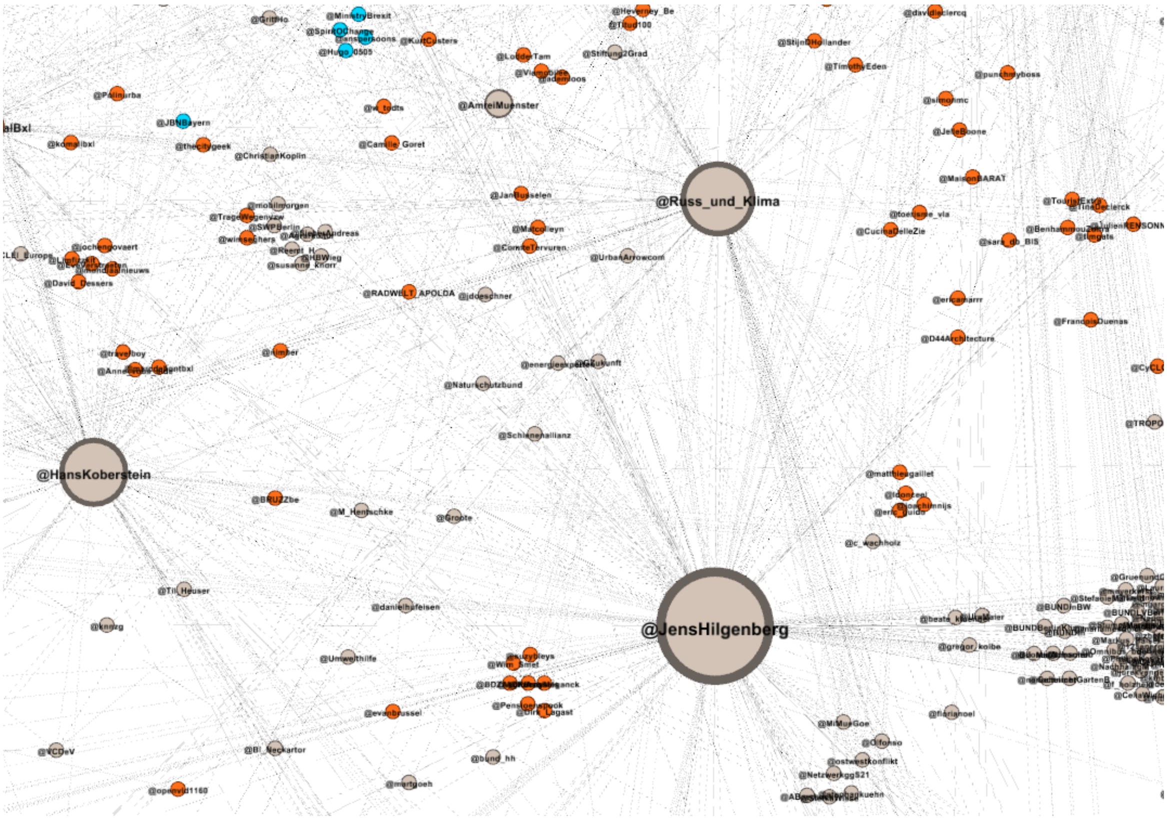

Figure 30: German accounts community of the hackAIR followers network

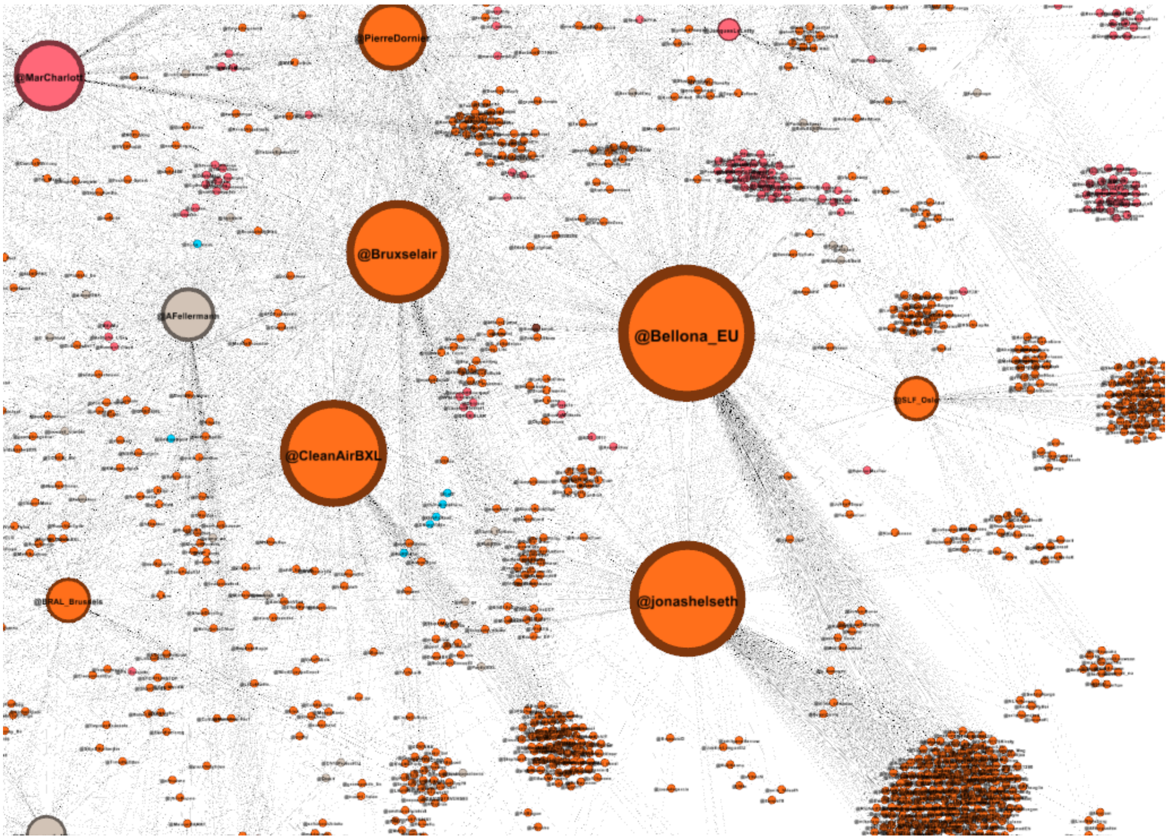

Figure 31: Belgian-Norwegian accounts community of the hackAIR followers network

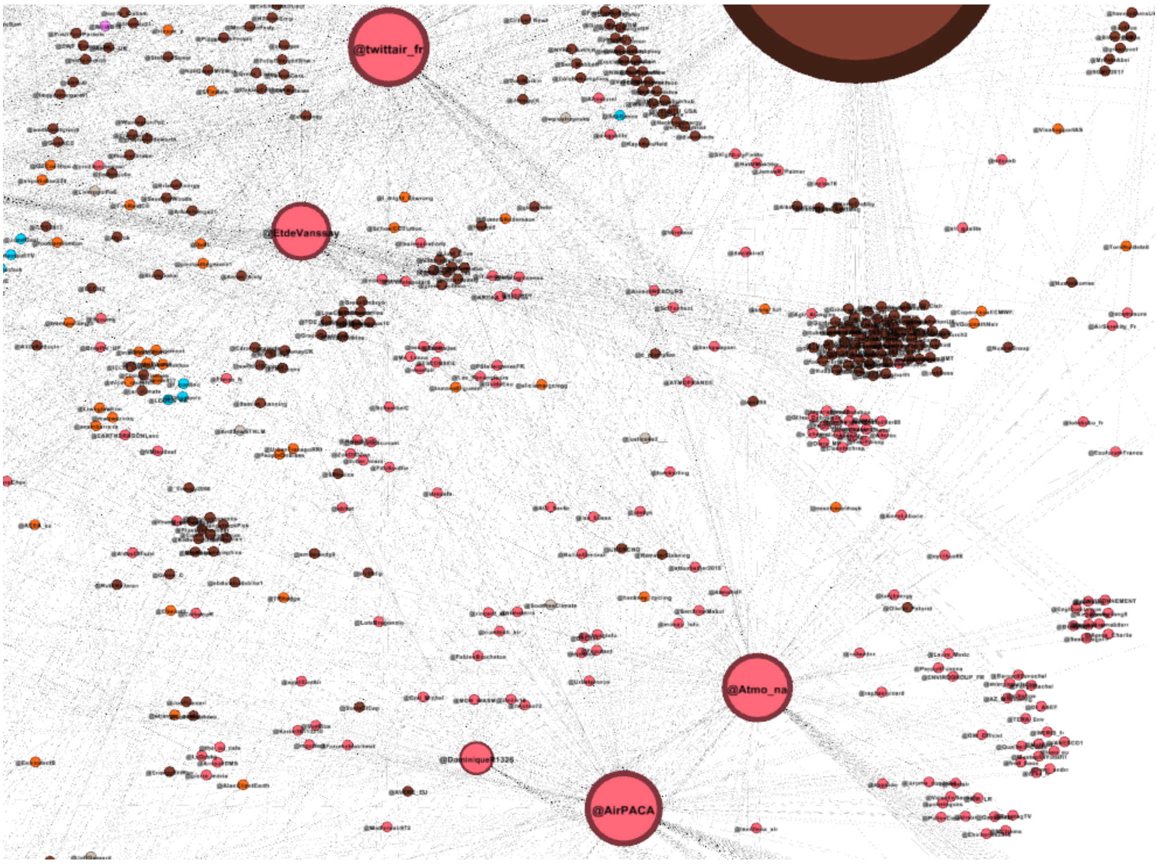

Figure 32: French accounts community of the hackAIR followers network

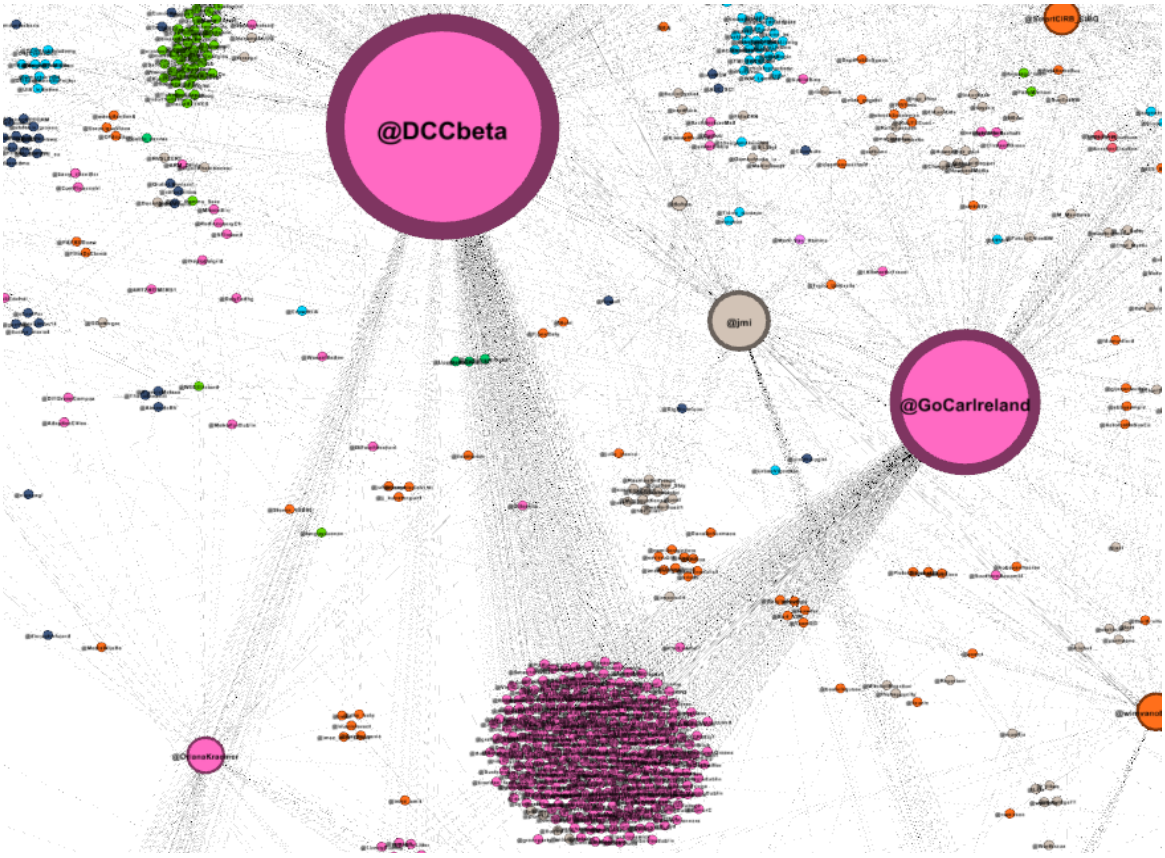

Figure 33: Irish accounts community of the hackAIR followers network

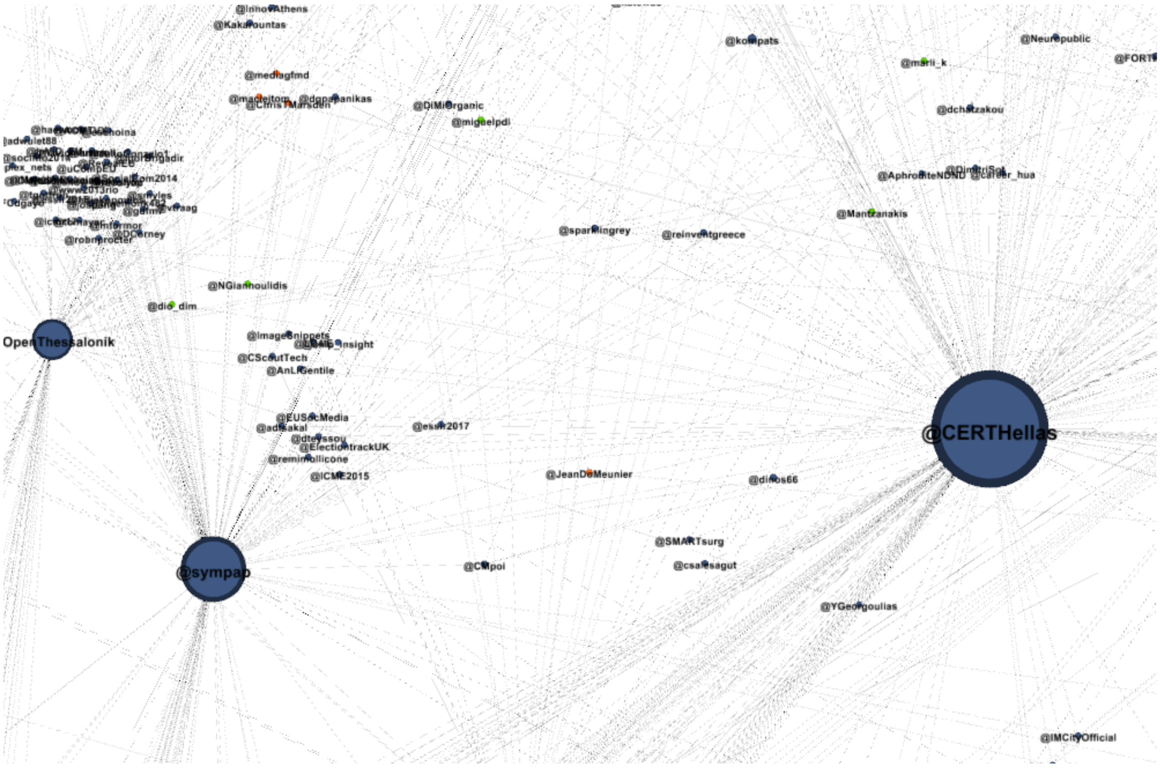

Figure 34: Greek accounts community of the hackAIR followers network

Table 12: Keywords used for hackAIR-relevant account discovery

Language	Keywords
English	air pollution, atmosphere pollution, particulates, particulate matter, car pollution, vehicle pollution, power plant pollution, fuel combustion pollution, burning coal, industrial pollution, pesticide pollution, vehicle emissions, coal pollution, Great Smog, sulfur dioxide pollution, carbon monoxide pollution, ground level ozone, air carcinogen, air pesticides, nitrogen dioxide pollution, co2 pollution, co2 emissions, pm10 pollution, pm2.5 pollution, no2 pollution, so2 pollution, air quality index, ozone pollution, waste incinerators, stratospheric ozone depletion, coarse particles, fine particles, nitrogen oxides, emission factors, air pollutant, ambient air pollution, unhealthy air, hazardous Air, urban air pollution, asthma air pollution, bronchitis air pollution, emphysema air pollution, chronic bronchitis air pollution, chronic obstructive pulmonary disease air pollution, respiratory illness air pollution, respiratory problems air pollution, respiratory disease air pollution, health air pollution, breathing problems air, cough air pollution, health problems air pollution, against air pollution, antipollution, anti-air pollution, stop air pollution, clean atmosphere, breathe air pollution, air pollution movement, prevent air pollution, smog pollution, ozone pollution, haze pollution, burning coal pollution, air pollution asthma, bronchitis air pollution, Germany air pollution, Netherlands air pollution, France air pollution, Norway air pollution, Belgium air pollution, air pollution alerts, air pollution alert, air pollution forecast, air pollution forecast, air pollution prediction, air pollution index, air pollution alert, factory emissions, air pollution observatory, pm10, pm25, co2, no2, so2, pm10 index, pm2.5 index, PM2.5 concentration, PM10 concentration, real time air pollution, air pollution map, air pollution activism, air quality, airpollution, airquality, polluted air, noxious gases, exhaust fumes, polluted atmosphere, air pollutants, car emissions, smog, antipollution, air particles, carbon dioxide pollution, o3 pollution, coarse particles, fine particles, nitrogen oxide pollution, toxic air, filthy air, unhealthy air, toxic smog, toxic emissions

German	Luftverschmutzung, Atmosphäre Umweltverschmutzung, Partikel, Feinstaub, Autoabgase, Kraftwerk Verschmutzung, Kohleverbrennung, brennende Kohle, Industrielle Verschmutzung, Fahrzeugemissionen, Kohleverschmutzung, Verschmutzungsrisiko, Fahrzeug Emissionen, Großer Smog, Kohleverschmutzung, Benzin Verschmutzung, Kohle Verschmutzung, Große Smog, Schwefeldioxidbelastung, Kohlenstoffmonoxid, Bodennahes Ozon, Luft Karzinogen, Stickstoffdioxidbelastung, Bleibelastung, CO2 Belastung, CO2 Emissionen, PM10 Belastung, PM2.5 Verschmutzung, no2 Verschmutzung, SO2 Verschmutzung, reaktives Gas, Luftqualitätsindex, Ozonbelastung, Müllverbrennungsanlagen, luftgetragenen Blei, Ozon in der Stratosphäre depleters, grobe Partikel, Feinpartikel, Stickoxide, Persistent freie Radikale, Toxische Metalle, Husten Luftverschmutzung, atmen Luftverschmutzung, Emissionsfaktoren, Luftschadstoff, Luftverschmutzung, ungesunde Luft, Gefährliche Luft, städtische Luftverschmutzung, Asthma Luftverschmutzung, Bronchitis Luftverschmutzung, Emphysem Luftverschmutzung, chronische Bronchitis Luftverschmutzung, chronisch obstruktive Lungenerkrankung Luftverschmutzung, Erkrankungen der Atemwege Luftverschmutzung, Atemprobleme Luftverschmutzung, Atem diseaseair Verschmutzung, Gesundheit Luftverschmutzung, Probleme mit der Atmung Luft, Husten Verschmutzung, Gesundheitsprobleme Umweltverschmutzung, gegen Luftverschmutzung, Ozonverschmutzung, Kohleverbrennung Luftverschmutzung, feinstaubwarnung, atmen Verschmutzung, Feinstaubalarm, verhindern die Luftverschmutzung, Smog Umweltverschmutzung, ozon Verschmutzung, Dunst Verschmutzung, Himmel Verschmutzung, Smog Dunst, Verbrennung von Kohle Verschmutzung, Vorhersage Luftverschmutzung, Kritische Luftbelastung, Luftverschmutzungsindex, Deutschland Luftverschmutzung, Niederlande Luftverschmutzung, Frankreich Luftverschmutzung, Norwegen Luftverschmutzung, Belgien Luftverschmutzung, Luftverschmutzung Warnungen, Luftverschmutzung Alarm, Luftverschmutzung Prognose, kritische Luftverschmutzung, polution Prognose, PM2.5 Index, Schadstoffindex, PM2.5 Konzentration, Luftschadstoff, Luft Messstation, PM10 Konzentration, Fabrikemissionen, Luftverschmutzung Observatorium, Luftverschmutzung Index, Luftqualität, Atmosphärische Luftbelastung, Kraftwerksemissionen, Kohlenstoffmonoxidbelastung, Feinstaubbelastungen, Feinstaubbelastung
French	la pollution de l'air, la pollution de l'atmosphère, particules, affaire particulière, pollution automobile, la pollution des centrales électriques, la pollution de la combustion de combustibles, charbon brûlant, pollution industrielle, risque de pollution, les émissions des véhicules, la pollution des carburants, menaces de pollution, la pollution de l'essence, la pollution du charbon, la pollution de l'anhydride sulfureux, la pollution de l'oxyde de canbon, L'ozone au niveau du sol, carcinogène de l'air, pesticides aériens, pollution par le dioxyde d'azote, , pollution par le plomb, pollution de co2, Emissions de CO2, pollution pm10, pollution PM2.5, pollution no2, pollution so2, gaz réactif, l'indice de qualité de l'air, pollution par l'ozone, incinérateurs de déchets, plomb d'origine atmosphérique, destructeurs d'ozone stratosphérique, des particules grossières, Particules fines, chlorofluorocarbone, oxydes d'azote, radicaux libres persistants, Les métaux toxiques, polluants radioactifs, peroxyacétyle nitrate, Les facteurs d'émission, polluant de l'air, la pollution de l'air ambiant, air malsain, Air dangereux, la pollution de l'air urbain, stagnation Air, pollution de l'air de l'asthme, pollution de l'air de la bronchite, pollution de l'air de l'emphysème, pollution de l'air de la bronchite chronique, pollution chronique de l'air de la maladie pulmonaire obstructive, pollution de l'air de maladie respiratoire, problèmes respiratoires pollution de l'air, la pollution des voies respiratoires diseaseair, pollution de l'air sur la santé, problèmes respiratoires air, la pollution de la toux, des problèmes de santé pollution, contre la pollution, qualité de l'air, lutte contre la pollution de l'air, arrêter la

	pollution de l'air, atmosphère propre, respirer la pollution, le mouvement de la pollution, prévenir la pollution de l'air, pollution par le smog, pollution ozon, pollution de la brume, pollution du ciel, smog brume, brûler la pollution du charbon, la pollution climatique, la pollution de l'asthme, la pollution de la bronchite, pollution de l'air allemagne, alertes de pollution de l'air, Alerte pollution de l'air, Prévisions de la pollution de l'air, pollution de l'air critique, Prévisions polution, la prédiction de la pollution, indice de pollution, Alerte pollution, polluant de l'air, pollution du ciel, pollution par le smog, pollution industrielle, les émissions des usines, pollution de l'air observatoire, indice de pollution de l'air
Dutch	luchtvervuiling, atmosfeer vervuiling, deeltjes, fijnstof, vervuiling door auto's, energiecentrale verontreiniging, verbranding van de brandstof vervuiling, brandend kool, industriële vervuiling, vervuiling pesticide, gevaar voor vervuiling, emissies van voertuigen, brandstof vervuiling, bedreigingen vervuiling, benzine vervuiling, kolen vervuiling, grote Smog, zwaveldioxideverontreiniging, canbon koolmonoxide vervuiling, air kankerverwekkend, air pesticiden, stikstofdioxide vervuiling, co2 vervuiling, CO2 uitstoot, PM10 vervuiling, PM2.5 vervuiling, no2 vervuiling, SO2 vervuiling, luchtkwaliteit, ozonverontreiniging, afvalverbrandingsinstallaties, Stratospheric ozonaantastende stoffen, grove deeltjes, Fijne deeltjes, stikstofoxiden, emissiefactoren, luchtverontreinigende stoffen, luchtverontreiniging, ongezonde lucht, gevaarlijke Air, luchtverontreiniging in de steden, astma luchtvervuiling, bronchitis luchtvervuiling, emfyseem luchtvervuiling, chronische bronchitis luchtvervuiling, chronische obstructieve longziekte luchtverontreiniging, ziekte van de luchtwegen luchtvervuiling, ademhalingsproblemen luchtvervuiling, gezondheid luchtvervuiling, ademhalingsproblemen lucht, gezondheidsproblemen vervuiling, anti luchtvervuiling, stoppen met luchtvervuiling, schone sfeer, ademen vervuiling, beweging vervuiling, voorkomen dat de luchtvervuiling, ozon vervuiling, hemel vervuiling, verbranding van kolen vervuiling, astma vervuiling, bronchitis vervuiling, duitsland luchtvervuiling, Nederland luchtvervuiling, france luchtvervuiling, noorwegen luchtvervuiling, België luchtvervuiling, alerts luchtvervuiling, alert luchtvervuiling, prognose luchtvervuiling, kritieke luchtvervuiling, index vervuiling, fabriek uitstoot, observatorium luchtvervuiling, luchtvervuiling index, atmosfeervervuiling, zwevende deeltjes, energiecentrale vervuiling, elektriciteitscentrale vervuiling, vervuiling door energiecentrales, brandstof verbranding vervuiling, vervuiling door verbranding van fossiele brandstoffen, steenkool verbranding, steenkoolverbranding, overbemesting, Voertuigemissies, voertuigemissie, brandstofvervuiling, vervuilingsrisico, vervuilings risico, steenkool vervuiling, smoggevaar, smogalarm, zwaveldioxidevervuiling, Koolstofmonoxide vervuiling, koolmonoxide vervuiling, koolstofmonoxidevervuiling, luchtvervuilende stoffen, kankerverwekkende stoffen, kankerverwekkende lucht, pesticiden in de lucht, luchtkwaliteitindex, luchtkwaliteit index, ozonvervuiling, stratosferische ozonaantastene stoffen, ozonaantastende stoffen, luchtvervuilende stoffen, luchtvervuiling, gevaarlijke lucht, luchtvervuiling in de stad, longziekte luchtvervuiling, chronische obstructieve longziekte luchtverontreiniging, anti luchtvervuiling, stop luchtvervuiling, ademen luchtvervuiling, luchtvervuiling voorkomen, ozonvervuiling, vervuiling door verbranding fossiele brandstoffen, steenkoolverbranding vervuiling, Frankrijk luchtvervuiling, waarschuwing luchtvervuiling, luchtvervuiling meldingen, voorspelling luchtvervuiling, luchtvervuilingsindex, pm2, 5 concentratie, pm10 concentratie, actuele luchtvervuiling, luchtkwaliteitskaart, activisme luchtvervuiling
Norwegian	luftforurensing, atmosfærisk forurensning, partikler, svevestøv, bileksos, industriell forurensning, forurensning pga brenning av drivstoff, brenning av kull, industriell forurensning, forurensning pga plantevernmiddel, biologisk forurensing, forurensningsfare,

	utslipp fra kjøretøy, forurensning pgs drivstoff , forurensningstrussel, forurensning pga bensin, kull forurensning, svoveldioksid forurensning, karbonmonoksid forurensning, bakkenært ozon, kreftfremkallende stoffer i luften, plantevernmidler i luften, nitrogendioksid forurensning, bly forurensning, co2 forurensning, co2 utslipp, PM10 forurensning, PM2.5 forurensning, no2 forurensning, SO2 forurensning, reaktiv gass, luftkvalitetsindeks, ozon forurensning, avfallsforbrenningsanlegg, grove partikler, fine partikler, klorfluorkarbon, nitrogenoksider, persistente frie radikaler, giftige metaller, Radioaktiv forurensing, Peroxyacetylnitrat, utslippsfaktorer, luftforurensning, ambient luftforurensning, usunn luft, farlig luft, urban luftforurensning, stagnasjon av luft, astma luftforurensning, bronkitt luftforurensning, emfysem luftforurensning, kronisk bronkitt luftforurensning, kronisk obstruktiv lungesykdom luftforurensning, luftveissykdom luftforurensning, luftveisproblemer luftforurensning, luftveissykdom luftforurensning, helse luftforurensning, pusteproblemer luft, hoste forurensning, helseproblemer forurensning, mot forurensning, mot forurensning, mot luftforurensning, stopp luftforurensning, ren atmosfære, å puste forurensning, forurensning bevegelse, forhindre luftforurensning, forurensning pga smog , forurensning pga ozon, forurensning pga dis, forurensning av himmelen, smog dis, forurensning pga kullfyring, klimaforurensning, luftforurensning varslinger, luftforurensning varsling, luftforurensningsprognose, kritisk luftforurensning, forurensningsprognose, forurensningsforutsigelse, forurensningsindeks, forurensningsvarsling, himmel luftforurensning , forurensning pga smog , industriell forurensning, utslipp fra fabrikken, luftforurensningsobservatoriet, luftforurensningsindeks, pm10 indeks, pm2.5 indeks, pm2.5 konsentrasjon, pm10 konsentrasjon, luftkvalitetsindeks observasjon, sanntids luftforurensning, luftforurensningskart, luftforurensningsaktivisme, luftkvalitet

Table 13: Influential accounts based on their incoming degree

Username	Link	Incoming degree
@BatesJenny	www.twitter.com/BatesJenny	1171
@networknicola	www.twitter.com/networknicola	1156
@kresin	www.twitter.com/kresin	1041
@JoBurgess11	www.twitter.com/JoBurgess11	951
@sfnohohon	www.twitter.com/sfnohohon	905
@SFotonium	www.twitter.com/SFotonium	896
@maponi	www.twitter.com/maponi	749
@miskaknapek	www.twitter.com/miskaknapek	702
@RomainWeikmans	www.twitter.com/RomainWeikmans	666
@yann_drareg	www.twitter.com/yann_drareg	661
@stuartberry1	www.twitter.com/stuartberry1	614
@andythegreenie	www.twitter.com/andythegreenie	590
@Jan_sagt	www.twitter.com/Jan_sagt	571
@jonashelseth	www.twitter.com/jonashelseth	556
@Eddyca1	www.twitter.com/Eddyca1	543

@markjleach	www.twitter.com/markjleach	516
@bertrandwert	www.twitter.com/bertrandwert	509
@mcmahoneth	www.twitter.com/mcmahoneth	471
@MoniqueCalisti	www.twitter.com/MoniqueCalisti	461
@DVRansbeeck	www.twitter.com/DVRansbeeck	436
@JessCappadonna	www.twitter.com/JessCappadonna	431
@casperkoomen	www.twitter.com/casperkoomen	397
@basvanheur	www.twitter.com/basvanheur	385
@SarahWest_SEI	www.twitter.com/SarahWest_SEI	381
@FelicettoMassa	www.twitter.com/FelicettoMassa	376
@Sina__Hn	www.twitter.com/Sina__Hn	343
@mking007	www.twitter.com/mking007	333
@MarCharlott	www.twitter.com/MarCharlott	327
@f_pilla	www.twitter.com/f_pilla	316
@PierreDornier	www.twitter.com/PierreDornier	315
@JensHilgenberg	www.twitter.com/JensHilgenberg	292
@pallavipnt	www.twitter.com/pallavipnt	288
@Ju_Fink	www.twitter.com/Ju_Fink	286
@IvonneLeenen	www.twitter.com/IvonneLeenen	261
@langleydew	www.twitter.com/langleydew	258
@WoonTan	www.twitter.com/WoonTan	256
@phebedudek	www.twitter.com/phebedudek	243
@LucIntPanis	www.twitter.com/LucIntPanis	242
@AFellermann	www.twitter.com/AFellermann	239
@tcolehunter	www.twitter.com/tcolehunter	238
@smkraus_	www.twitter.com/smkraus_	229
@lindyfrey	www.twitter.com/lindyfrey	206
@EtdeVanssay	www.twitter.com/EtdeVanssay	199
@AlisonDyke_SEI	www.twitter.com/AlisonDyke_SEI	196
@natamyk	www.twitter.com/natamyk	194
@busybee_75	www.twitter.com/busybee_75	192
@BurfordBen	www.twitter.com/BurfordBen	190
@RachelWrangham	www.twitter.com/RachelWrangham	186
@RachelWrangham	www.twitter.com/RachelWrangham	186
@fpedrera	www.twitter.com/fpedrera	157
@HansKoberstein	www.twitter.com/HansKoberstein	154

@mikesara444	www.twitter.com/mikesara444	152
@sympap	www.twitter.com/sympap	145
@saskiacoulson	www.twitter.com/saskiacoulson	137
@FabiBenavente	www.twitter.com/FabiBenavente	117
@KatrienCJSM	www.twitter.com/KatrienCJSM	103
@skontopo	www.twitter.com/skontopo	101
@BatesJenny	www.twitter.com/BatesJenny	1171

Table 14: Most influential users that do not follow hackAIR account

Username	Link	Followers	# hackAIR followers it follows
@ShaunFrankson	www.twitter.com/ShaunFrankson	300,995	18
@tveitdal	www.twitter.com/tveitdal	229,127	22
@natalieben	www.twitter.com/natalieben	142,594	26
@Alex_Verbeek	www.twitter.com/Alex_Verbeek	132,483	17
@randal_olson	www.twitter.com/randal_olson	110,471	26
@climatehawk1	www.twitter.com/climatehawk1	74,996	25
@arikring	www.twitter.com/arikring	69,220	19
@JasonLRobinson	www.twitter.com/JasonLRobinson	64,362	19
@BradZarnett	www.twitter.com/BradZarnett	52,404	12
@Jackthelad1947	www.twitter.com/Jackthelad1947	30,099	15
@09Clive	www.twitter.com/09Clive	29,202	24
@blindspotting	www.twitter.com/blindspotting	26,116	26
@SustMeme	www.twitter.com/SustMeme	22,070	12
@RichardMcLellan	www.twitter.com/RichardMcLellan	18,838	10
@WYeates	www.twitter.com/WYeates	17,362	23
@DrJulieRiggs	www.twitter.com/DrJulieRiggs	12,915	23
@NicolasLoubet	www.twitter.com/NicolasLoubet	12,484	21
@TessyBritton	www.twitter.com/TessyBritton	10,682	13
@B_Moorhouse	www.twitter.com/B_Moorhouse	9,129	17
@GEO_Jan	www.twitter.com/GEO_Jan	8,670	11
@dr_rick	www.twitter.com/dr_rick	7,042	11
@Steven_Ramage	www.twitter.com/Steven_Ramage	5,718	14
@Lea_Shanley	www.twitter.com/Lea_Shanley	3,667	15
@IAQsusan	www.twitter.com/IAQsusan	3,032	25

@mwt2008	www.twitter.com/mwt2008	2,963	36
@KarlGeissler	www.twitter.com/KarlGeissler	2,047	13
@ShellyMBoulder	www.twitter.com/ShellyMBoulder	1,853	15
@susana_hgs	www.twitter.com/susana_hgs	1,651	21
@marinavance	www.twitter.com/marinavance	1,576	22
@Andrew9Boswell	www.twitter.com/Andrew9Boswell	1,537	11
@Lorelei350	www.twitter.com/Lorelei350	1,529	13
@DrTomasS	www.twitter.com/DrTomasS	1,528	22
@nyrojasr	www.twitter.com/nyrojasr	1,438	24
@AnAirThatKills	www.twitter.com/AnAirThatKills	1,420	54
@parthaabosu	www.twitter.com/parthaabosu	1,381	14
@jenniferbmuller	www.twitter.com/jenniferbmuller	1,341	11
@JoeriThijs	www.twitter.com/JoeriThijs	1,273	14
@maritvp	www.twitter.com/maritvp	1,235	14
@DrGaryHaq	www.twitter.com/DrGaryHaq	1,230	16
@PavlosDoikos	www.twitter.com/PavlosDoikos	1,221	14
@cwiedinm	www.twitter.com/cwiedinm	1,182	19
@delgerzul	www.twitter.com/delgerzul	1,152	17
@vfmcneill	www.twitter.com/vfmcneill	1,109	14
@JunoDee	www.twitter.com/JunoDee	1,045	11

References

- Aiello, L., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Goker, A., Kompatsiaris, I., Jaimes, A. (2013). Sensing trending topics in Twitter. Transactions on Multimedia, 15(6), 1268-1282.
- Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, P10008.
- Fan, W., & Gordon, M. (2014). The power of social media analytics. *Communications of the ACM, 57*(6), 74-81.
- Haynes, J., & Perisic, I. (2009, June). Mapping search relevance to social networks. In Proceedings of the 3rd Workshop on Social Network Mining and Analysis (p. 2). ACM.
- Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PloS one, 9(6), e98679.
- Kanungsukkasem, N., & Leelanupab, T. (2016). Power of crowdsourcing in Twitter to find similar/related users. 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 1-6, IEEE.
- Mejova, Y., Weber, I., & Macy, M. (2015). *Twitter: A Digital Socioscope.* Cambridge University Press.
- Papadopoulos, S., Kompatsiaris, Y., Vakali, A., & Spyridonos, P (2012). Community detection in social media: performance and application considerations. Data Mining and Knowledge Discovery 24(3), pp. 515-554, Springer.
- Schinas, M., Papadopoulos, S., Apostolidis, L., Kompatsiaris, Y., & Mitkas, P. (2017). Open-source monitoring, search and analytics over social media. *Proceedings of the Internet Science Conference (INSCI 2017).* Springer.