FINAL SOCIAL NETWORK ANALYSIS TRUST & PEOPLE SEARCH TECHNIQUES

SmartH2O

Project FP7-ICT-619172

Deliverable D4.5 WP4

Deliverable
Version 1.1 – 30 May 2017
Document. ref.: D45.POLIMI.WP4

Programme Name:	ICT
Project Number:	619172
Project Title:	SmartH2O
Partners:	Coordinator: SUPSI Contractors: POLIMI, UoM, SETMOB, EIPCM, TWUL, SES, MOONSUB, UPV, EMIVASA
Document Number:	smarth2o. D4.5.POLIMI.WP4.V1.1
Work-Package:	WP4
Deliverable Type:	Document
Contractual Date of Delivery:	30 September 2016
Actual Date of Delivery:	30 May 2017
Title of Document:	Final social network analysis trust & people search techniques
Author(s):	Piero Fraternali, Chiara Pasini, Jasminko Novak, Mark Melenhorst, Isabel Micheel, Luigi Caldararu, Alessandro Facchini, Mattia Brusamento
Approval of this report	Submitted for approval to EC
Summary of this report:	The SmartH2O approach to social network analysis and people search techniques
History:	see version history table
Keyword List:	social network analysis, Twitter, crawling, information campaigns
Availability	This report is public

This work is partially funded by the EU under grant ICT-FP7-619172

Disclaimer

This document contains confidential information in the form of the SmartH2O project findings, work and products and its use is strictly regulated by the SmartH2O Consortium Agreement and by Contract no. FP7- ICT-619172.

Neither the SmartH2O Consortium nor any of its officers, employees or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein.

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7-ICT-2013-11) under grant agreement n° 619172.

The contents of this document are the sole responsibility of the SmartH2O consortium and can in no way be taken to reflect the views of the European Union.

Document History

Version	Date	Reason	Revised by
0.1	20/08/2016	DDP ready	M. Melenhorst P. Fraternali I. Micheel
0.2	05/09/2016	Thames Water application case described	M. Melenhorst I. Micheel
0.3	12/09/2016	Set up of the Valencian case study	M. Melenhorst I. Micheel M. Brusamento
0.4	16/09/2016	Results of the Valencian case study	M. Melenhorst I. Micheel M. Brusamento
0.5	20/09/2016	Comparative analysis	M. Melenhorst I. Micheel
0.6	20/09/2016	Description of the social analytics software modules	C. Pasini, P Fraternali
0.7	23/09/2016	Updated SmartH2O Social Network Crawler and Data Analyzer Updated application to Spanish Case study	S. Dumitru L. Caldararu
1.0	26/09/2016	Revision of the lessons learned and conclusions. Executive summary Final format check	M. Melenhorst, I. Micheel, J. Novak, A.E. Rizzoli
1.1	30/05/2017	Revisions after project final review	C.Pasini, J. Novak, M. Melenhorst, A.E. Rizzoli

Table of Contents

1 INTRODUCTION 3

2 ANALYSING COMMUNITY ROLES 4

2.1.1 Relationship-based roles 4
2.1.2 Behaviour-based roles 4
2.2 BEHAVIOURAL METRICS AND INFLUENCER ANALYSIS 5

3 SMARTH2O APPROACH TO BEHAVIOURAL ROLE ANALYSIS 7

3.1 HYPOTHESIZED BEHAVIOURAL ROLES 7
3.2 DEFINITION OF BEHAVIOURAL DIMENSIONS AND METRICS 7
3.3 EXPERIMENTAL EVALUATION OF THE SMARTH2O APPROACH IN THAMES WATER DATASET 10
3.3.1 Dataset construction and pre-processing 10
3.3.2 Procedure 10
3.3.3 Results for the Thames Water dataset 11
3.4 COMMENTS 15

4 SMARTH2O SOCIAL NETWORK CRAWLER AND DATA ANALYZER 17

4.1 DATA MODEL OF THE SOCIAL NETWORK CRAWLER AND DATA ANALYZER 18
4.2 METRICS OF THE SOCIAL NETWORK CRAWLER AND DATA ANALYZER 19

5 APPLICATION OF THE SMARTH2O METHOD FOR BEHAVIOURAL ROLE ANALYSIS TO THE SPANISH CASE STUDY 23

5.1 DATASET CONSTRUCTION 23
5.2 PROCEDURE 23
5.3 RESULTS 24
5.3.1 Clustering results 24
5.3.2 Analysis and interpretation of behavioural roles 26

6 LESSONS LEARNT FROM COMPARATIVE ANALYSIS OF TWO APPLICATION CASES 30

6.1 COMPARATIVE ANALYSIS 30
6.2 LESSONS LEARNT AND RECOMMENDATIONS 32

7 REFERENCES 36

APPENDIX A 38

Executive Summary

This deliverable reports on the SmartH2O approach to social network analysis and people search techniques, which in the Annex I of the Project Contract is described as follows:

"Final social network analysis trust & people search techniques: This work package contains the final results of the analysis of the social networks and user's behaviour. It describes the methods and tools developed/adopted in SmartH2O to identify, analyse, and model the social network of water consumers, based on the experience and data of the two case studies. It describes the user community with the identified structure of relationship and roles. Finally, it presents the lessons learnt and recommendations providing a comparative analysis of the two water user communities and identification of the most effective leverage to stimulate water use change."

The deliverable describes the developed SmartH2O method for the analysis of behavioural roles from social networks (Twitter) that identifies relevant behavioural roles and user groups for supporting targeted communication campaigns for utilities (e.g. for promoting water saving), The method is realized as a software component integrated into the SmartH2O platform, including an easy-to-use user interface (social dashboard), that allows utilities to identify relevant groups of users in social networks (e.g. Twitter) who can best be targeted in communication campaigns as different types users most likely to propagate campaign messages through the social network, such that these will be noticed by others.

The deliverable starts with a critical synthesis of existing behaviour-based and relationship-based approaches to the identification of such users most relevant with respect to the chosen SmartH2O approach (**Section 2**). The SmartH2O method for social network analysis that goes beyond existing approaches with the goal of being directly applicable to support multi-faceted communication campaigns is described in **Section 3**. Behavioural metrics are devised that help to identify different behavioural roles from Twitter interactions as indicators that allow predicting which users can support the utilities' promotion campaigns. The internal validity of the method is tested by applying it to the social network of the Thames Water Twitter account (data sample of 6,3 million tweets). The results identified five behavioural roles: authorities, disseminators, creators, interactionals, and observers, with distinctive behavioural patterns regarding suitability for supporting communication campaigns.

The technical realization of this method as a software component of the SmartH2O platform – the SmartH2O Social Network Crawler and Data Analyzer - is described in **Section 4**. This includes its architecture, the data model and an easy-to-use user interface (the Social Dashboard), that allows interactive exploration of the results of the behavioural analysis in an intuitive, easily accessible way. To verify its robustness the method has been applied to another application case, the SmartH2O Spanish case study in Valencia (**Section 5**.) This case allows the verification of the external validity of the method, as it addresses a completely different geographical, cultural and language community than in the first case. An alternative crawling approach for the construction of the data set (5 million tweets) has also been applied, covering a more varied range of sample users. The results reveal the same types of clusters, behavioural roles and very similar associated behavioural patterns as in the first application. This confirms the validity of the developed method and its applicability to targeted communication campaigns in social networks.

The deliverable concludes with a comparative analysis of the two application cases (Thames Water and Valencia) and presents lessons learnt for multifaceted targeted communication campaigns and recommendations for supporting water saving (**Section 6**). The presented results suggest that the proposed method for analysis of behavioural roles in Twitter can identify a set of user types (and users) that can support information dissemination in Twitter campaigns in a multi-faceted, targeted manner. The consistent findings between the cases suggest that the introduced behavioural dimensions and metrics are suitable for capturing behavioural patterns that can support targeted information dissemination. The identified roles and their distribution in the two cases demonstrate that a much larger portion of active users, suitable for supporting Twitter-based communication campaigns, can be identified with this method than with other approaches. Accordingly, the described findings provide direct

support for designing targeted communication campaigns by water utilities within the SmartH2O context. Finally, based on the lessons learnt from the SmartH2O experience recommendations for effective means of supporting water saving are presented.

1 Introduction

For the water utilities in the SmartH2O consortium, social media form an increasingly important part of their marketing efforts to reach out to their customers and to extend the interaction between the utility and the customers. The use of social media channels can support these strategic communication tools, contributing to campaigns for e.g. saving water, for assessing general customer satisfaction, or for evaluating the impact of a consumer awareness initiative.

The effectiveness of such campaigns is dependent on social media users to propagate messages to their friends and followers. The analysis of user behaviour on social networks with respect to the propagation of messages can improve the campaign results of water utilities, contributing to the business goals of utilities.

While substantial research has studied the overall dynamics and behavioural patterns of Twitter communication, little work has investigated how such analysis could support communication campaigns in a goal-directed way. The SmartH2O approach is to identify and characterize different types of behavioural roles of social network (Twitter in the present effort) users that can be used to support communication campaigns (that aim at attracting user attention on specific issues) in a targeted manner. The implemented method is based on a set of dimensions that can capture specific behavioural roles of Twitter users from their communicational patterns and the results of its application to an experimental dataset from a concrete application scenario (environmental awareness).

We first assess related work in this area, with regard to community roles analyses and behavioural metrics. Subsequently we describe the approach integrate within SmartH2O, and then report on the experimental evaluation we undertook first on a dataset composed of Thames Water followers, and then, to validate the findings, on a dataset constructed from accounts and tweets from the Valencia area. The deliverable concludes with a discussion on how the results of such an application can be used to support communication campaigns in a multi-faceted way, addressing different user types.

2 Analysing community roles

This section overviews the different possible approaches and sets the ground for the specific approach chosen in SmartH2O.

2.1.1 Relationship-based roles

Relationship-based or structural roles are derived from social network analysis applied to a graph, where users are represented as nodes connected with edges that represent either some stated connection (e.g. friends, followers) or online interactions. The roles are derived from different centrality measures (e.g. in-/outdegree, betweenness), often combined with community detection, and reflect intra- and inter-community positions in the network [Guimera & Nunes, 2005].

An example are the roles proposed in [Scripps et al., 2007] that consider the number of communities to which a node is connected and its degree centrality: e.g. the Ambassadors being characterized by a high degree and connected to different communities, Bridges defined as being also connected to different communities but having a low degree. A similar approach in [Fagnan et al., 2014] differentiated between roles global or local to the community (or both) and identified the following role types: Extrovert, the Mediator, the Peripheral, the Outlier and the Principal (the latter also divided into Community leaders and Global principles). In [Beguerisse-Díaz et al., 2014] Twitter users were distinguished with respect to their role in information propagation into: References that have large audience of followers but who themselves follow a small number of users; Engaged Leaders who have large number of followers but few friends and often interact; Mediators who interact with both leading categories and the listeners, Diversified Listeners who have few followers and follow different types of accounts suggesting thus a diversity in their interests and information sources following and Listeners, also with few followers who themselves follow mostly References and are considered passive recipients. References are typically institutional accounts, popular personalities or information sources, Engaged Leaders entail institutional accounts but also personal accounts and Mediators include often journalists and reporters. Similarly, this approach is concerned primarily with the numbers of followers and friends of each user, neglecting other important aspects, such as the actual communication patterns. Other structural approaches infer the roles using only network centrality measures, such as [Tyshchuk et al., 2013] that discovered three different types of leaders with respect to their role in Twitter during emergency cases: the Diffuser, the Gatekeeper and the Information broker.

2.1.2 Behaviour-based roles

Despite their contribution of important knowledge about the structure of a network and the possibility to use associated metrics, such structural analyses tell us little about the actual context and content of the relations. For example, a structurally derived role of a “broker” (a person linking two otherwise disconnected groups of people) although theoretically important, does not effectively distinguish between individuals who actually engage in brokering behaviour from the ones who don’t [Gleave et al., 2009].

Analysing the behavioural patterns that users exhibit in their communication and interaction with others is likely to better serve that purpose. In fact, part of the difficulty in defining the very concept of role becomes obvious when user roles are identified based on behavioural patterns. A segmentation of users based on their behaviour is often referred to as user types instead of roles. For example, in [Brandtzaeg & Heim, 2004] the ‘Lurker’ is part of a user typology, while in [Golder & Donath, 2004] it is a considered a role. However, a role is by definition a concept associated also with behavioural attributes while an online social network can be seen as a large community where users are related to each other and interact. Furthermore, behavioural patterns entail in many cases information about the interactions. Measuring, for example, the retweet ratio of a user conveys information about both the user’s

relationships and interactions within the broader community of Twitter.

A number of different behaviour-based user roles or typologies have been identified so far in Twitter interactions. In [Tinati et al., 2012] users in Twitter were categorised into specific roles based on their communication behaviour, aiming mainly at identifying users who are potentially producers or distributers of valuable content. The non-mutually exclusive roles include the Idea Starters, who create unique content on social media; the Amplifiers, who accumulate thoughts and share them to their large network of connections; Curators as users who have a broader context and tend to validate or challenge specific views thus influencing the way the conversation is shaped; Commentators who contribute their own insights without seeking leadership and prestige; and Viewers as the least active category that includes rather passive users. Though in itself a plausible model, grounded in insights from marketing practice, the metrics for determining these roles are rather heuristic: e.g. Idea Starters are defined as users whose original tweets are in total retweeted more than a defined threshold, while Amplifiers are those whose number of tweets is greater than their retweets multiplied by first-in-chain retweets [Tinati et al., 2012]. This approach only considered the retweeting activity of users and although its findings seem very plausible, behaviour of users in social networks is likely to be better reflected by taking into account more varied aspects. Several other elements like the number of own unique tweets, replies, followers or friends can also be of great importance in the identification of behavioural roles in online interaction.

While [Maulana & Tjen, 2013] examined both Facebook and Twitter and found 6 clusters of users, the roles they distinguish are hardly applicable to our application scenario of communication campaigns: they differentiate between e.g. Angels, as individuals interested in business networking, sharing knowledge and advice; Active and Passive Learners, users who learn actively and share their knowledge vs. those who gain knowledge but rarely share theirs etc. Similarly, other research in dynamics of online communities such as Usenet [Golder & Donath, 2004] has identified roles that can be found in every virtual community (e.g. Newbies, Celebrities, Lurkers), which also have rather limited applicability to our scenario (except for Celebrities, as users who contribute actively to the community and are popular through their frequent activity). Somewhat closer are the roles of Leaders and Motivators identified in Usenet by [Pal & Counts, 2011]. In [Brandtzaeg & Heim, 2011], a questionnaire-based study regarding user types in Social Networking Services (SNS) in general, conceptualized user types by their participation objective (whether recreational or informational) and the level of participation. It revealed 5 distinct clusters of which Socialisers, Debaters and Actives may have some relevance for supporting communication campaigns, but whose behavioural properties are not modelled by behavioural metrics that would allow an assessment of their relevance in this context.

2.2 Behavioural metrics and influencer analysis

Besides the described approaches to identifying different behaviour roles and types of users, different approaches have proposed various behavioural metrics for examining and modelling user behaviour without classifying it into specific roles or aimed at identifying just one type of most relevant users (e.g. so-called influencers).

Influential users are frequently defined as those that have the potential to achieve the highest information diffusion. Such approaches often attempt to identify influencers based on the ratio of the number of followers to friends [Weng et al., 2010] (assuming that a large number of followers represents a large potential audience for information diffusion) or by considering graph-based social network properties (e.g. in-degree, retweet and mention influence) [Cha et al., 2010] as reflecting reachable audience. However, previous research on link farming and “social capitalists” ([Ghosh et al., 2010; Dugué & Perez, 2014] shows that large numbers of followers can be relatively easily gained by reciprocal following as a common practice to increase social capital - without actual interest in each other’s content and accordingly with low information propagation.

Other actions, such as the act of retweeting, have been shown to carry a stronger indication of topical relevance [Welch et al., 2011]. Similarly, comparisons of different measures of influences (in-degree, retweet and mentions) have found that users who have high in-degree

are not necessarily influential in terms of spawning retweets (i.e. of triggering information diffusion) [Cha et al., 2010]. In addition to such limitations of follower-based approaches, we are not only interested to identify the so-called influential users that clearly are simply prominent or prestigious in the community. In terms of effect on information spreading, a user with a large audience who almost never disseminates content of others (e.g. of a source of a communication campaign) is equally irrelevant for this application case, as is a user with only few followers. Such a conclusion is also supported by [Wang et al. 2012] that showed how actual information spreaders are very unlikely to be the important persons in the egocentric networks, and even unlikely to be important persons globally. The more suitable users for our purpose should be not only able but also willing to disseminate information from others. This doesn't exclude also involving the notion of influential users but it requires a different definition and different behavioural metrics to identify them.

One possibility is to consider metrics for identifying the most authoritative users proposed in [Pal & Counts, 2011], such as the Retweet Impact (entailing how often the user was retweeted and by how many unique users), the Mention Impact (measuring the mentions that user receives for reasons other than responses to him mentioning others) and the Signal Strength (measuring how much the user contributes with original content to the topic). Similarly, the indicators used to measure role compositions and identify behavioural patterns in online communities proposed in [Rowe & Alani, 2012] could be adapted to our context. This approach was applied to an online blog, but their methodology could be adapted to Twitter. In order to model and measure user behaviour, they used six dimensions with corresponding metrics. Most relevant to our goal include: **Engagement**, regarding the proportion of accounts that a user has replied to; **Popularity**, regarding the proportion of users that replied to the user; **Contribution**, regarding the ratio of threads the user created to the total number of threads; **Initiation**, regarding the number of threads the user started; and **Content Quality**, regarding the points awarded to each user about their answers.

As we have seen, modelling user behaviour such that it allows the identification of different roles is inherently linked to deciding what aspects of user behaviour are to be captured. Moreover, most frequent approaches to supporting information dissemination on Twitter focus on one type of most prominent users, the influencers and consider the retweet rate and follower links as the most important metrics for their identification. In contrast, our goal of addressing different types of users in a multi-faceted way requires a number of different behavioural dimensions and associated metrics to be considered, to capture different behavioural patterns. Existing models of behaviour roles (user typologies), while providing some points of departure for modelling behavioural attributes, do not derive the user roles with respect to our application scenario. In order to identify behavioural roles/types of Twitter users that exhibit behavioural patterns most relevant for supporting multi-faceted communication campaigns, we needed to develop an approach that integrates the described insights and adapts them into a solution tailored to this specific purpose.

3 SmartH2O approach to behavioural role analysis

The goal of the proposed method for the analysis of behavioural roles is to enable the identification of users that could most effectively spread the content of a given Twitter campaign in their network and increase the likeliness of other users noticing it. In developing such a method of analysis we proceed as follows:

i) we hypothesize potentially relevant behavioural types for this purpose,

ii) define behavioural dimensions and metrics to characterize them appropriately and

iii) apply these to a concrete case to verify that they can yield behavioural types suitable for supporting goal-directed information spreading among different types of users.

This approach and the developed method are described in the next sections of this chapter.

3.1 Hypothesized behavioural roles

Considering results of existing work on behavioural roles reviewed before and our stated goal, we identified a preliminary set of types of behavioural roles that 1) we could expect to exist in Twitter and 2) would be suitable for supporting Twitter campaigns.

The most obvious candidate for spreading content of others are users who are more willing to retweet than to create own content, so the role of Amplifier, Diffuser or Disseminator (similar to those discussed in Chapter 2) should be naturally present. Users that mostly use Twitter as a discussion forum, debating or discussing over topics and news (also in smaller groups), would likely be frequently referencing other users as part of the discussion, which creates notifications that trigger user attention on those messages (by making them stand out from the general flow of tweets). Thus, a type of Debater or Commentator would hold a place in our hypothetical typology. Users holding seemingly important positions due to a large reach could also be potentially influential disseminators, if they can be triggered to disseminate one's content. Accounts belonging to famous personalities, to local or international organisations or to users that gained some prestige or fame exclusively in Twitter and social media are likely candidates for this role. Accordingly, grouped together, one would likely characterize them as Celebrities or Influentials. Naturally, depending on the granularity one is trying to achieve, these user types could be certainly split into more refined versions. Finally, no user typology would be complete without the usual Lurkers (or Viewers or Passive Learners) describing the users that are rather inactive or may passively consume information without leaving much trace to be identified. These are just as important, as they are typically the largest part of online networks and communities. Based on such a hypothetical model of behavioural roles and their purpose in our context, we proceeded to define behavioural dimensions that could capture the behavioural patterns of such user types.

3.2 Definition of behavioural dimensions and metrics

Defining specific behavioural dimensions and associated metrics for identifying such behaviour roles requires the interpretation of possible meaning of user actions on Twitter [Honeycutt & Herring, 2009]. User actions are commonly encoded with typical metrics of Twitter activity, such as number of tweets, mentions, followers etc. For the definition and calculation of such metrics and behavioural dimensions we consider three different kinds of tweets, which derive mostly from three different types of actions: (new) content creation, content dissemination and conversation. This conceptualization is similar to [Pal & Counts, 2011] that categorized the tweets of each user into following categories: original tweets (OT), conversational tweets (CT) and repeated tweets (RT). We keep the same categorization but refine the classes to contain conceptually the most common cases. In particular, under repeated tweets we include three different cases of tweets produced by others and forwarded

by the user to his/her audience: any retweet that is identified by the Twitter metadata (as made through the built-in retweet button and thus we call this native retweet); tweets that are manually made and are identified by the prefix RT followed by @username or starting as "@username, just like those replacing RT with MT indicating modified retweet (we call such cases manual retweet) and finally, the quoted statuses that are identified also through the tweet metadata (called here quote retweets). In all these cases a user is disseminating another tweet and therefore we name them Dissemination tweets (DT), instead of RT (see Table 1). The definitions of CT and OT match the ones of [Pal & Counts, 2011]. Conversational tweets (CT) are tweets directed at another user and are identified by the fact that they are denoted by the use of @username token preceding the text. Original tweets (OT) are those written by the user (creating own content) in such a way that doesn't fall under the other two categories.

The reason for establishing this classification is that the above categories of tweets have different roles and textual format. The user contributes new content using OT or disseminates content with DT while CT conveying direct communication between him/her and other Twitter members. In addition, there is a difference in terms of the visibility of the content as both OT and RT will appear in the timelines of user's followers whereas CT have the peculiarity that they appear only on timelines of those users who are common followers of the sender and the recipient. Thus, it is important to consider these differences. Especially in the exploration of user interests, the differentiation of all these tweet types illustrates the variability of reading and writing aspects. Accordingly, we define seven dimensions that could capture and model different types of user behaviour that can support multifaceted communication campaigns in Twitter, as follows:

1. **Activity:** the frequency of user's activity is a critical measure, indicating how much a user participates in Twitter. We consider this to be reflected in the average number of tweets per day. A square root is applied to reduce the right skewness of the data.

$$Activity = \sqrt{\frac{total\ tweets\ acquired}{covered\ period\ (days)}}$$

2. **Content Contribution:** the proportion of new content a user contributes can differentiate users who contribute (more) own content from those who tend to propagate the content of others. This dimension reflects the portion of tweets with new content (OT) the user contributes out of all the tweets he produces, escalated by the logarithm of original tweets.

$$Content\ contribution = \frac{OT}{Total\ tweets} \log(OT + 1)$$

3. **Dissemination:** This dimension aims at expressing to what extent the user acts as a disseminator of the content of others. Dissemination is a crucial factor for supporting the propagation of messages of a communication campaign. It is calculated as the ratio of dissemination tweets (DT) to the total number of user's tweets, multiplied by the logarithm of the total dissemination tweets, which reflects the scale of the number of dissemination tweets posted by the user (see Table 1). The retweet history is a good indicator for retweet prediction only if a user is a frequent retweeter [Wang et al., 2012], and thus we include the scale of the retweets.

$$Dissemination = \frac{DT}{Total\ tweets} \log(DT + 1)$$

4. **Direct Interaction:** This dimension represents how much a user engages in direct communication with other users. The key indicator here is the number of replies of the user either to other tweets or directly to a user. The number of distinct users a given user engaged with plays also a role, as it makes a difference if someone

addresses a few same people or many different ones. It is estimated as the proportion of user's conversational tweets (CT) to the total number of tweets of the user, multiplied by the logarithm of the distinct users the user has replied to. This indicates what part of the user's activity is devoted to direct interaction and at what scale.

$$Direct\ interaction = \frac{CT}{Total\ tweets} \log(RT\ engagements + 1)$$

5. **Retweet impact:** this dimension aims to reflect the extent to which a user's tweets are retweeted, whilst including the impact that his content has on others. This content may be mostly own content (OT) but also content of others that he further spreads. This dimension concerns only a subset of tweets: tweets that are meant to be visible in the tweet stream (not CT) and contain solely or partly own content from the user, i.e. they are either OT or dissemination tweets which the user rephrased or in which s/he included own comments. These tweets are from now on denoted in this work as Twuc; Tweets with (visible) user content.

$$Retweet\ impact = \frac{T_{wuc}\ Retweeted}{T_{wuc}} \log(RT\ Engagements + 1)$$

Dissemination tweets are discriminated accordingly in Table 1. The use of logarithm is on the one hand to scale the total and to dampen the effect of same people often retweeting the user.

Table 1. Discrimination between dissemination tweets.

Type of tweet	Form	Own content
RTn (native)		N
RTm (manual)	MT@	Y
	"@ text" comment	Y
	Comment RT@	Y
	"@ text"	N
	RT@	N
QT (quote)		Y

6. **Popularity:** Popularity can illustrate how much attention a user can get regarding his/her potential audience. This dimension is calculated by two factors, a ratio of the number of followers to the sum of followers and friends, and the number of times a user has been added to other users' curated lists. The ratio tends to 1 as the user is being followed by more people than he follows, to 0 in the opposite case and to 0.5 when the number of followers and friends are close and thus it can be used as an indicator of a user's popularity. We consider lists as a secondary indicator of popularity. Users create lists to group others together under their perception that they share some property in common, which can vary (profession, type, common interest etc.). Thus the popularity of the enlisted people lies in the fact that they have likely drawn the user's attention strongly enough to make the effort to list them. However, we don't give this a high weight as we can't take for granted that the users actually follow the news of the lists or that any list assignment is a positive thing. Thus the logarithm of the number of lists is used to contain its effect on the dimension.

$$Popularity = \frac{Followers}{Followers + friends} \log(listed\ count + 1)$$

7. **Attention Triggering:** This dimension aims at describing a user's impact on drawing attention of other users to a message by explicitly mentioning them in his tweets: using @username and thus causing notifications to be displayed to these users. Since users are much more likely to notice tweets from such notifications than those in their general tweet stream, this measure is an important indicator of a user's ability to trigger other users' attention. It is calculated as the ratio of the number of times the user mentioned some other user to the number of all tweets, adjusted by the scale level of the number of distinct users mentioned (log). Total tweets here exclude the native retweets, as in this case the user only contributes in the spreading of someone else's tweet with mention(s). Mentioning someone either by replying to his tweets or by including him in new ones is likely to trigger the user to perform an action, such as replying or at least reading the tweet. The logarithm of the number of distinct users mentioned allows us to differentiate the case of few users mentioned very often from the case of mentioning many different users a few times each.

$$Attention\ triggering = \frac{Mentions}{Total\ tweets\ excl.\ RT_n} \log(distinct\ users + 1)$$

After modelling the user data with the described dimensions, each user's behavioural patterns are encoded with a corresponding behavioural vector. Groups of users with similar behavioural patterns can then be identified by applying a clustering method. By examining the properties of the determined clusters the behavioural roles characterizing a given group of users can be identified and representative users to be targeted in a communication campaign selected. An experimental application of this method is discussed in the sub section.

3.3 Experimental evaluation of the SmartH2O approach in Thames Water dataset

3.3.1 Dataset construction and pre-processing

The dataset used was based on the followers of the Twitter account of a large European water utility involved in the project, exemplifying a potential application context in which such a company is interested in performing a communication campaign on Twitter by pushing information to its followers. Another assumption is that users following the Twitter account of such a company do so because they are interested in different topics related to the company and its operations (e.g. water, water consumption, company news & service information, water alerts etc.). The company in question had approximately 20'000 followers when the crawling process was conducted. Out of these, only accounts with public timelines that were located in the company's water supply area were selected. This was inferred either from statements in their profiles or, where available, based on their geographical coordinates. After the crawling, further filtering was conducted by excluding protected users, users with very low activity (less than 11 tweets in total) and users with no posts in the current year (considered as inactive). Two more users were excluded as assumed bot accounts due to excessively high daily activity of over 400 tweets per day. The final data sample resulted in 4.437 users (out of the initial 4.961 users). For each user, their timeline was crawled using Twitter's Search API (allowing the fetching of up to 3.200 tweets per user). In total, 6,3 million (6.314.906) tweets were gathered.

3.3.2 Procedure

Modelling users by defined behavioural metrics

The dataset of selected users was modelled along the dimensions described in Section 3.2,

by calculating the defined dimension metrics for each user. Potential correlations in the dimensions were tested with the Spearman correlation coefficient (as a robust measure not requiring normal distribution in the data). We found a possible but not clear correlation between two pairs of dimensions: RT Impact and Popularity (0.68); Direct Interaction and Attention Triggering (0.62). Given the inconclusive correlation values, we decided to keep both pairs of dimensions and to see what the clustering results may suggest in terms of their possible differentiation value.

Clustering by shared behavioural patterns

Having described each user's behavioural pattern with the introduced behavioural dimensions and associated metrics, an unsupervised clustering was performed on the user's behavioural vectors in order to group users sharing similar behavioural patterns. Before clustering, the data was normalized to prevent dimensions with larger ranges (e.g. Activity, Attention Triggering) to outweigh the rest. The clustering was performed by applying the K-means algorithm. To determine the optimal number of clusters we iteratively ran the algorithm increasing the number of clusters from k=2 to k=18 and repeated this procedure 22 times, recording each time the value of the Davies-Bouldin Index. After comparing the mean DB-Index for each k, the smallest index (DB=1.22) was established for k=5. Accordingly, a K-means clustering with k=5 was performed, obtaining five clusters representing groups of users with similar behavioural patterns.

3.3.3 Results for the Thames Water dataset

Clustering results

The values of individual dimensions for the centroids of the identified clusters are depicted in Figure 1, which indicates some obvious differences between the clusters. Cluster 0 (yellow line) has a very obvious peak in the Direct Interaction dimension and the highest value in Attention Triggering among all clusters, while a low mean in all other dimensions, especially in Dissemination, RT Impact and Popularity. This indicates a group of users who use Twitter as a medium of direct discussion in groups, i.e. a user type that we term "Interactionals".

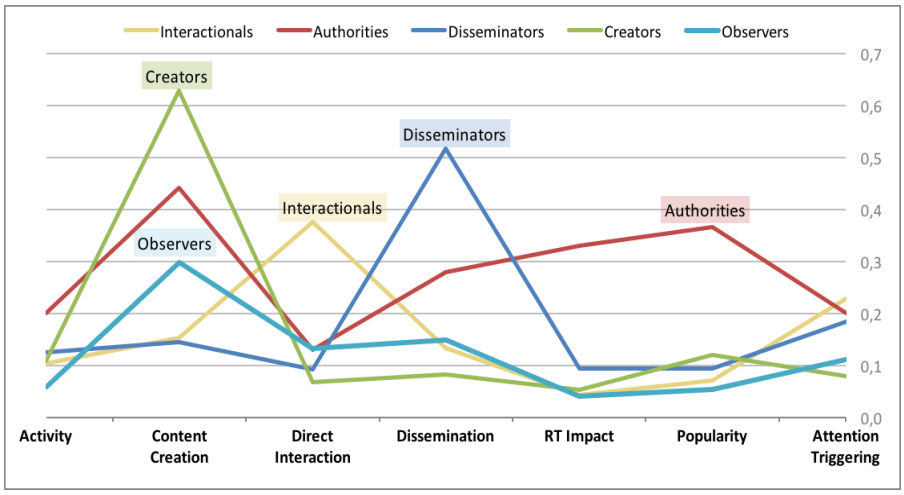


Figure 1. Values of behavioural dimensions of cluster centroids.

In contrast, Cluster 1 (red line) presents the highest average values in Activity, Popularity and RT Impact, with also high Content Creation and Dissemination dimensions, suggesting the most active and most noted group of users. While mentioning other users also seems to be frequent in this group (second highest value of Attention Triggering among all clusters), this seems to happen more in their tweets to all, than in direct communication with specific users.

The combination of the highest values in RT Impact and Popularity dimension combined with Attention Triggering suggests an influential user type that we term “Authorities”.

Cluster 2 (blue line) is characterized by the highest Dissemination among all clusters, with moderate Activity and very low Content Creation and Direct Interaction. It also exhibits the second highest average in RT Impact and Attention Triggering close to the highest value. This suggests a user type with high ability of propagating content of others, that we term “Disseminators”. Cluster 3 (green line) concentrates clearly its activity in Content Creation with very low levels in dissemination and direct discussions. These users mention other users the least of all clusters and their RT Impact is low. Popularity is slightly above the other clusters but much lower than that of the Authorities cluster. This indicates accounts with a higher number followers and list memberships than the average, but with too low impact and triggering to be relevant. We term this user type “creators”. Finally, cluster 4 (turquoise line) includes the least active users on average, who concentrate their little activity on Content Creation, have low Attention Triggering and present the lowest levels of Popularity and RT impact. Accordingly, this type of users we relate to the usual role of “observers” (or often called lurkers).

In Figure 2 the means, standard deviations, and distribution of the values are depicted for each dimension per behavioural cluster. As can be seen, the within-cluster variability of individual dimension values is comparable across clusters. The graphs also show clearly recognizable differences in mean values of individual dimensions for different clusters, pointing to their distinctive character (as already depicted in a condensed form in Figure 1).

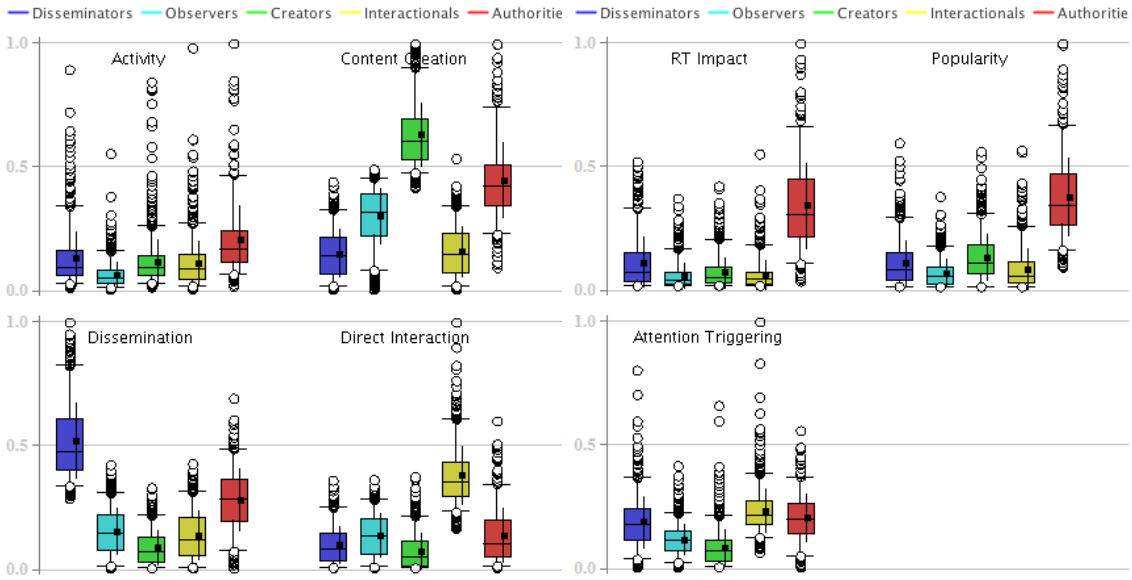


Figure 2. Distribution of values of the behavioural dimensions per cluster.

The distribution of users in the clusters is shown in Figure 3. It is no surprise that a big proportion of users act as Observers, while only a small group includes users with seemingly greatest impact (Authorities, 9%). But another 35% of users fall into two behavioural roles that also exhibit active behavioural patterns. The Disseminators present not only high Dissemination, but also information diffusion potential in terms of RT Impact. The Interactionals may not disseminate or information to large audiences, but they highly engage specific user groups and thus may activate them (Direct Interaction, Att. Triggering). Such evidence supports our approach to multi-faceted user roles for information spreading, beyond the usual influencer approaches: focusing only on the Authorities (typical “influencers”) would

neglect a big part of active users who can provide valuable support for the dissemination of communication campaigns.

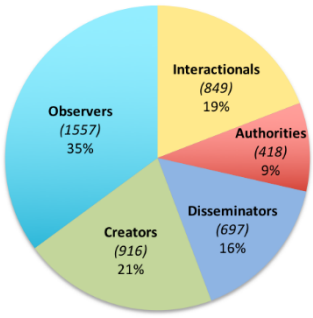


Figure 3. Distribution of users in clusters of behavioural roles.

This first analysis of the cluster characteristics based on the defined behavioural dimensions suggests the existence of behavioural roles similar to the ones we hypothesized; some of which may lend themselves well to supporting communication campaigns (e.g. Authorities, Disseminators). To better understand these findings, we next take a closer look at additional (meta)data characterizing the behavioural patterns of these groups of users.

Analysis and interpretation of behavioural roles

In order to shed more light on the behavioural patterns of the identified clusters, we analyze them with respect to additional data from the dataset, summarized in Table 2 and Table 3 (user and tweet statistics). We also examine users in the proximity of the cluster centroids (the 10 users closest to the centroids), to get some insight on typical users in a given behaviour role.

Interactionals: Users in this group tend to be more interested in direct communication and discussions. More than every second tweet (56.5%) in this group starts with the token @username, meaning that it is directed to a specific user and has less public visibility. Based on the metadata of the tweets, 46.7% of all tweets are replies to other statuses, and only 8.8% of them initiate a discussion with a user without a specific tweet preceding it. This suggests that many of these users use Twitter like a discussion forum. Users in this cluster often mention public or private services and comment on the services they are provided; giving either positive or negative feedback. Their posts do not have a significant retweet impact, as only 1 out of 10 tweets with own content is on average retweeted, although some of the users have a valuable number of followers and relatively high activity.

The greatest part of their activity are replies or directed messages of their initiation (56.5%). They write direct messages almost 3 times more than they create or disseminate content. The average activity is 2.1 tweets/day and the average number of followers 520.3. They rarely use media (9.5% of tweets) or URLs (14.5% of tweets) in their tweets but more often hashtags (21.1% of tweets). Involving these users in disseminating a campaign would have to be less media and more conversation-oriented. Examining the most representative users for this cluster (ten closest to the centroid), we find almost exclusively personal accounts of users who engage in discussions and/or contact brands through Twitter. Additionally, among the most popular and influential users (highest values for Popularity and RT Impact) we can find some journalists, commentators and few companies.

Authorities: This group of users presents the highest activity with 6.13 tweets/day. On average, every second tweet they create is retweeted around 5 times (2.635 RT engagements per T_{wuc}). This is reflected in highest average RT Impact and Popularity.

This group thus seems to get the most attention in terms of the dissemination of their content. They present high values in all dimensions but Direct Interaction. Almost half of their activity is about creating new content (OT ratio). They disseminate tweets of others less, but still at almost 1/3 of their activity (DT ratio). They address other users directly in almost 1 in 5 tweets (CT ratio), which is still notable in terms of potential for attention triggering. As expected, they have a big audience of followers (22573 on average), significantly larger than the average number of friends. They are listed very often by other users and 99% of them are at least in 5 lists. They include very often URLs and hashtags in their tweets (46.5% and 38.6%, respectively). They also share media, mostly images, but to a lower rate (18.3%). This indicates that they highly share information sources (thus likely also communication campaigns) and use hashtags to make their messages easily accessible to users out of their standard audience.

Table 2. User statistics per behavioural cluster.

	C0 Interact-ionals	**C1** Author-ities	**C2** Dissemi nators	**C3** Creators	**C4** Observe rs
Tweets /day	2.09	6.13	2.90	2.20	0.67
OT	391.53	1330.32	344.31	1193.36	365.16
CT	907.70	502.17	273.30	234.86	243.39
DT	306.39	841.50	959.11	209.22	184.74
OT ratio	0.220	0.502	0.210	0.764	0.477
CT ratio	0.589	0.178	0.157	0.116	0.266
DT ratio	0.191	0.321	0.633	0.120	0.257
Listed Count	13.58	241.84	18.32	22.65	6.62
Followers Count	520.3	22573.1	766.0	2837.3	326.8
Friends Count	749.8	6204.0	986.5	2171.2	552.6
Users replied in CT (distinct)	214.2	229.2	99.6	77.8	66.8
Users mentioned (dist.)	326.2	508.3	209.4	191.6	122.9
Engagements (RT/Fav) per T_{wuc}	0.551	4.214	0.925	0.383	0.435
T_{wuc} retweeted ratio	0.119	0.515	0.23	0.117	0.13
Followers/ (Friends+Followers)	0.311	0.704	0.341	0.412	0.308
Engagements (RT) per T_{wuc}	0.248	2.635	0.527	0.227	0.223

Examining the most representative users for this cluster (closest to the centroid), we find mostly institutional accounts: companies, non-profit organisations, local newspapers, business alliances and an independent university but also a politician and an environmental journalist. Among those with highest RT Impact, we find mostly popular newspaper accounts and local institutional accounts like the mayor and the fire brigade.

Disseminators: The principal characteristic of this group of users is that they tend to disseminate information much more than contributing own content or discussing with others. Despite the fact that they act mostly as amplifiers of the information posted by others, they present a moderate to high RT Impact, that is the second highest after the Authorities and almost double as high as of the three remaining groups. This is likely re-enforced by their Attention Triggering score, which is almost as high as of the Authorities. At least 1 out of 4 tweets they create is retweeted. These users are relatively active with an average of 2.9 tweets/day and they tend to follow others more than being followed themselves (average number of follower 766 and friends 986.5 (see Table 2). The latter indicates that they may be pursuing a strategy of acquiring many information sources from which they can retweet content to others. They present the highest use of media (23.7% of tweets) in comparison to other clusters and more than 1/3 of the tweets contain hashtags. This information can be used as a guide: campaigns targeting these users can be adapted to their preferences by e.g. containing media. Every third tweet contains URLs, which makes these users also

suitable for spreading links to outside information. Examining the most representative users for this cluster, we find personal accounts and a part of them seems to use Twitter in a profession-oriented manner.

Table 3. Tweet statistics of the behavioural clusters.

Tweets with:	**URLs**	**media**	**hashtags**
C0 Interactionals	0.145	0.095	0.211
C1 Authorities	0.465	0.183	0.386
C2 Disseminators	0.378	0.237	0.371
C3 Creators	0.480	0.089	0.326
C4 Observers	0.226	0.106	0.290

Creators: The main characteristic of this group is that the greatest part of their activity is concentrated on the creation of own content (on average 76.4% of user's activity is new content). The rest is divided almost equally between dissemination and direct messages. They have on average a regular but moderate activity (2.2 tweets/day). Additional inspection of the content of the Creators' tweets showed that almost every second tweet in this cluster contains a URL[1], while almost 1 of 3 tweets contain hashtags in this group whereas the use of media is not very popular and is the lowest in comparison to the other groups (8.9% of tweets). This seems to correspond to a previous study [Enge, 2014] showing that the use of images increases the likelihood of gaining retweets, since these users use very little media and score low in RT Impact. Despite the low RT Impact, they present a moderate to high Popularity, which corresponds to the large average number of followers (2837, see Table 2). This is likely a consequence of the "follow-to-be-followed" strategy that typically results in many reciprocal links with little retweeting effect [Dugué & Perez, 2014] (as the low Dissemination and RT Impact values testify).

Observers: These users have the lowest average activity with 0.67 tweets per day, the lowest Popularity and RT Impact average. When they tweet, they tend to create content twice as much as they disseminate tweets or engage in direct communication. The 90% of these users seem to follow more than being followed, with 2/3 of them having much less followers than they have friends (Table 2). They use to some extent URLs (22.6% of tweets) and a bit more often hashtags (29%) but rarely media (10% of tweets). Such data suggest that these users are mostly passive readers, consuming information, rather than creating or forwarding it to others, and that they use Twitter only occasionally.

3.4 Comments

This section has described the chosen SmartH2O approach to behavioural role analysis and the developed novel method to implement it, alongside with its application to an extensive dataset based on the Thames Water Twitter community. The results have confirmed the presence of the hypothesized behavioural roles, with the added value of being able to pinpoint a larger share of 'active' users than commonly present in influencer approaches.

The presented results suggest that the proposed method for analysis of behavioural roles in Twitter can identify a set of user types (and users) that can support information dissemination in Twitter campaigns in a multi-faceted, targeted manner. The experimental application to a dataset of 6,3 million tweets from a concrete application context (Twitter account of Thames Water, a large water utility interested in environmental awareness campaigns) has uncovered a set of user groups exhibiting behavioural roles suitable for this purpose. The identified roles

[1] Such a high value for URLs in tweets may indicate spammers who want to redirect the users to another page. An examination of a random sample of such users supports this assumption.

are similar to those initially hypothesized based on existing literature, but are more closely related to the specific needs of the stated application context. This suggests that the introduced behavioural dimensions and metrics are suitable for capturing behavioural patterns that can support targeted information dissemination.

In particular, the identified roles in the experimental dataset allow the identification of a larger portion of active users than in usual influencer approaches (44% users in active roles vs. 9% for influencers only) that could support communication campaigns.

The analysis of additional metadata (use of media, URLs, hashtags) indicated differences in behavioural patterns that can be used to tailor communication messages to suit the patterns of a given user type (behavioural role). A synthesis of the properties of the identified behavioural roles with the tweet metadata and the properties of typical users provides a concrete example of how multi-faceted communication campaigns (supporting dissemination among different types of users) could be performed in a targeted manner.

In section 5 the approach will be applied to another case (Aguas de Valencia, in Spain) in order to test whether the behavioural patterns and clusters of users are stable across different contexts, in different datasets.

Before that, Section 4 shows how the developed methods have been concretely embedded in the SmartH2O software architecture.

4 SmartH2O Social Network Crawler and Data Analyzer

After having been experimentally applied on the Thames Water Twitter community dataset, the described method has been implemented and extended into a fully-fledged software component for social network community roles analysis and integrated in the SmartH2O platform. The social analysis, trust and people search capability of SmartH2O is thus realized through a dedicated software component of the SmartH2O architecture, recalled in Figure 4, called **Social Network Crawler and Data Analyzer** (see deliverable D6.2 PLATFORM ARCHITECTURE AND DESIGN).

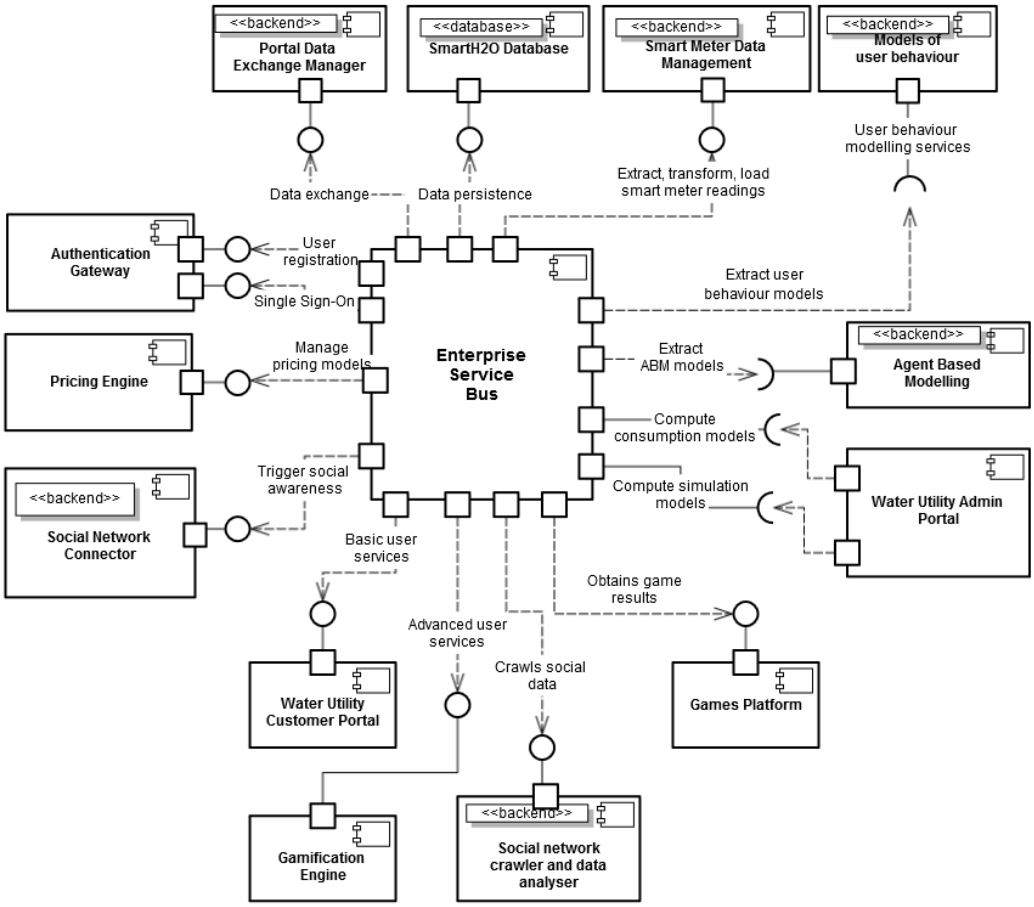


Figure 4. Architecture of SmartH2O (recalled from D6.2 PLATFORM ARCHITECTURE AND DESIGN).

The Social Network Crawler and Data Analyzer component allows the platform, where deemed appropriate by the water utility portal, to launch social data analysis experiments to identify relevant users and content in the area of sustainable water consumption. In the current implementation, this component supports the crawling of Twitter data in order to automatically find people and content relevant for a thematic area, such as water consumption.
Concretely, it performs the following functions:

- Crawling the content from a social platform APIs (in the present version, content comes from the twitter platform).

- Consolidating the crawled content into a temporary local database, for speeding up the analysis.
- Applying behavioural classification rules, in order to qualify people based on their influence status in the community.
- Formatting the results of the analysis to prepare them for visualization.
- Publishing the formatted analysis results into a Web application, part of the SmartH2O Admin Portal.

Figure 5 pictorially represent the internal organization of the Social Network Crawler and Data Analyzer component.

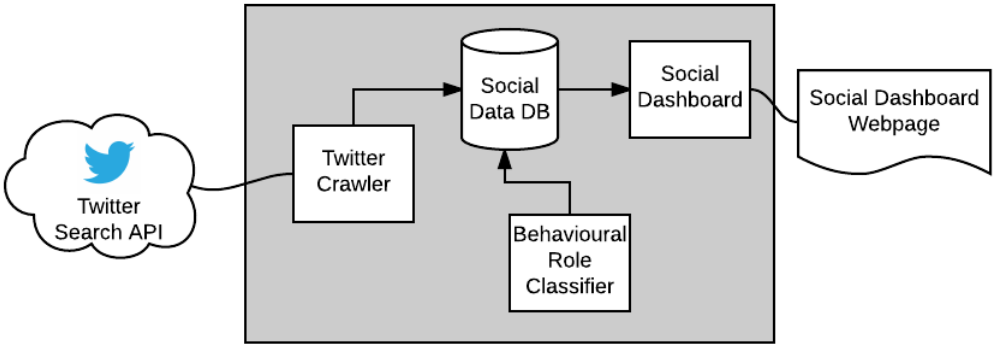


Figure 5. Architecture of the Social Network Crawler and Data Analyzer.

4.1 Data model of the Social Network Crawler and Data Analyzer

In order to proceed with the analysis, data from social networks are normalized according to the conceptual model illustrated in Figure 6.

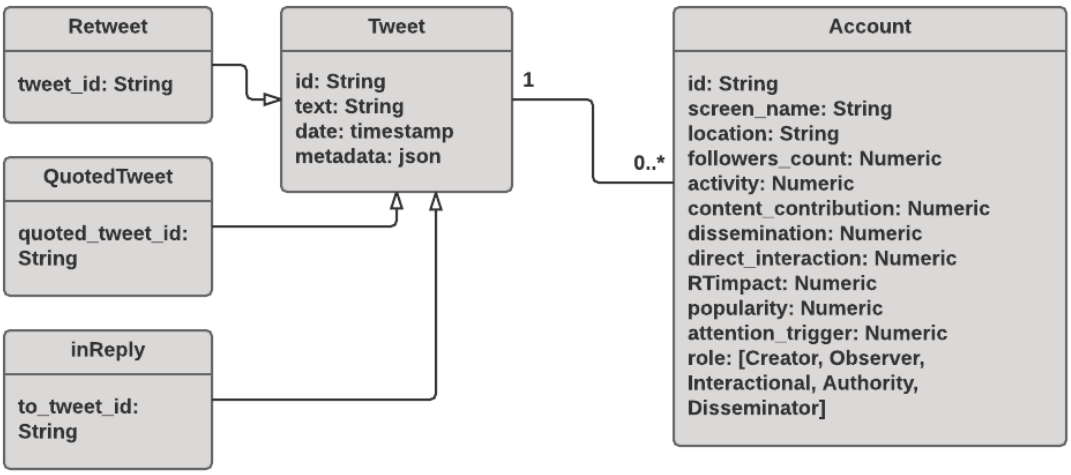


Figure 6. Conceptual model of the temporary database underlying the Social Network Crawler and Data Analyzer.

Content items are represented by the **Tweet** entity, which specializes into three sub-classes (**Retweets**, **QuotedTweets** and **inReply** tweets). They are associated with the **Account** entity, which represents the relevant properties of a social media user, including the profile

data and the calculated metric indices.

4.2 Metrics of the Social Network Crawler and Data Analyzer

The Social Network Crawler and Data Analyzer process the crawled content to compute the behavioural dimensions and metrics defined in Section 3.2, which are:

1. **Activity:** computed as the average number of tweets per day, square rooted to reduce the right skewness of the data.
2. **Content Contribution:** defined as the portion of original posts, i.e., Tweets with owned content, (OT) the user contributes out of all the tweets he produces, escalated by the logarithm of original tweets.
3. **Dissemination:** calculated as the ratio of dissemination tweets (DT) to the total number of user's tweets, multiplied by the logarithm of the total dissemination tweets, which reflects the scale of the number of dissemination tweets posted by the user.
4. **Direct Interaction:** estimated as the proportion of user's conversational tweets (CT) to the total number of tweets of the user, multiplied by the logarithm of the distinct users the user has replied to.
5. **Retweet impact:** proportion of the user's TWUC, Tweets with (visible) user content, that are retweeted.
6. **Popularity:** calculated as the product of two factors, a ratio of the number of followers to the sum of followers and friends, and the number of times a user has been added to other users' curated lists.
7. **Attention Triggering:** calculated as the ratio of the number of times the user mentioned some other user to the number of all tweets, adjusted by the scale level of the number of distinct users mentioned (log).

Figure 7 shows the main page of the Social Dashboard interface of the Social Network Crawler and Data Analyzer, where all behavioural dimensions and metrics are displayed.

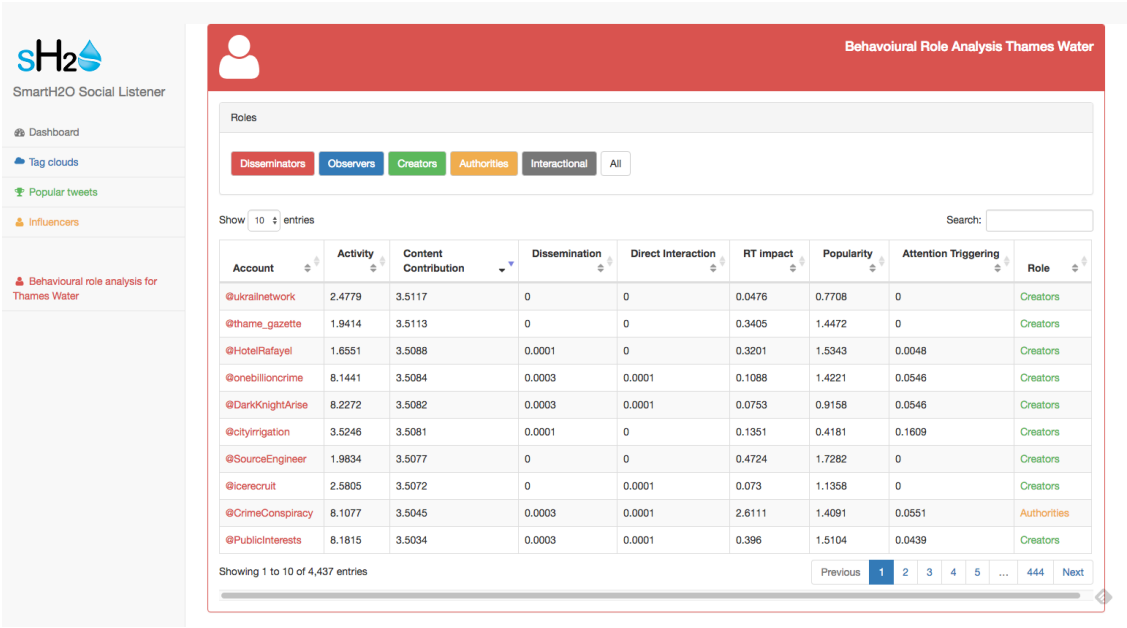


Figure 7. Behavioural dimensions and metrics displayed in the interface of the Social Network Crawler and Analyzer (a.k.a. Social Dashboard).

The described dimensions allow computing the user's behavioural patterns and identifying users with similar features. The Network Crawler and Analyzer apply a supplied clustering method and identify the relevant user groups, which are then displayed in the Social Dashboard (see Figure 8). Presently, the K-means clustering method is applied, but the Network Crawler and Analyzer has an open-ended architecture and allows the plug-in of any desirable data normalization and clustering technique.

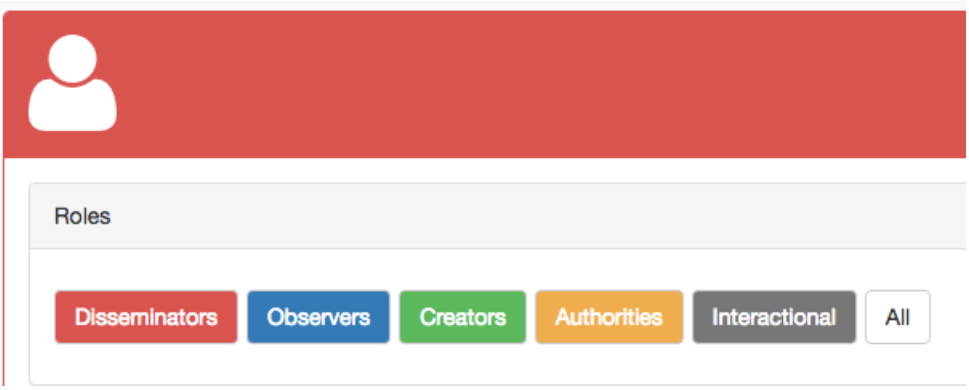


Figure 8. The identified clusters are represented as Roles in the Social Dashboard.

The Social Dashboard interface allows the operator to inspect users by role, so to be able to address them with communication campaigns best suited to their social profile.

Account	Activity	Content Contribution	Dissemination	Direct Interaction	RT impact	Popularity	Attention Triggering	Role
@CrimeConspiracy	8.1077	3.5045	0.0003	0.0001	2.6111	1.4091	0.0551	Authorities
@carrolltrust	8.3601	3.5024	0.0004	0.0001	2.3853	2.0674	0.0764	Authorities
@BillionDollarID	8.165	3.5016	0.0003	0.0001	2.9414	1.4769	0.0549	Authorities
@ConstructionEnq	2.5886	3.4974	0.0003	0.0032	2.7676	2.4911	0.0019	Authorities

Figure 9. Inspecting the features of users of group "Authorities" in London case study.

Besides the visual inspection and targeting of users based on their roles, the Social Dashboard allows the operator to have a panoramic view of the content at the base of the analysis.

Figure 10 shows the interface for content browsing. The upper part of the interface summarizes the most relevant terms appearing in the online discussions, divided into terms that appear in the post regular text and special-syntax terms (hashtags in the case of Twitter).

The lower part of the interface represents the salient social media posts and the influential users.

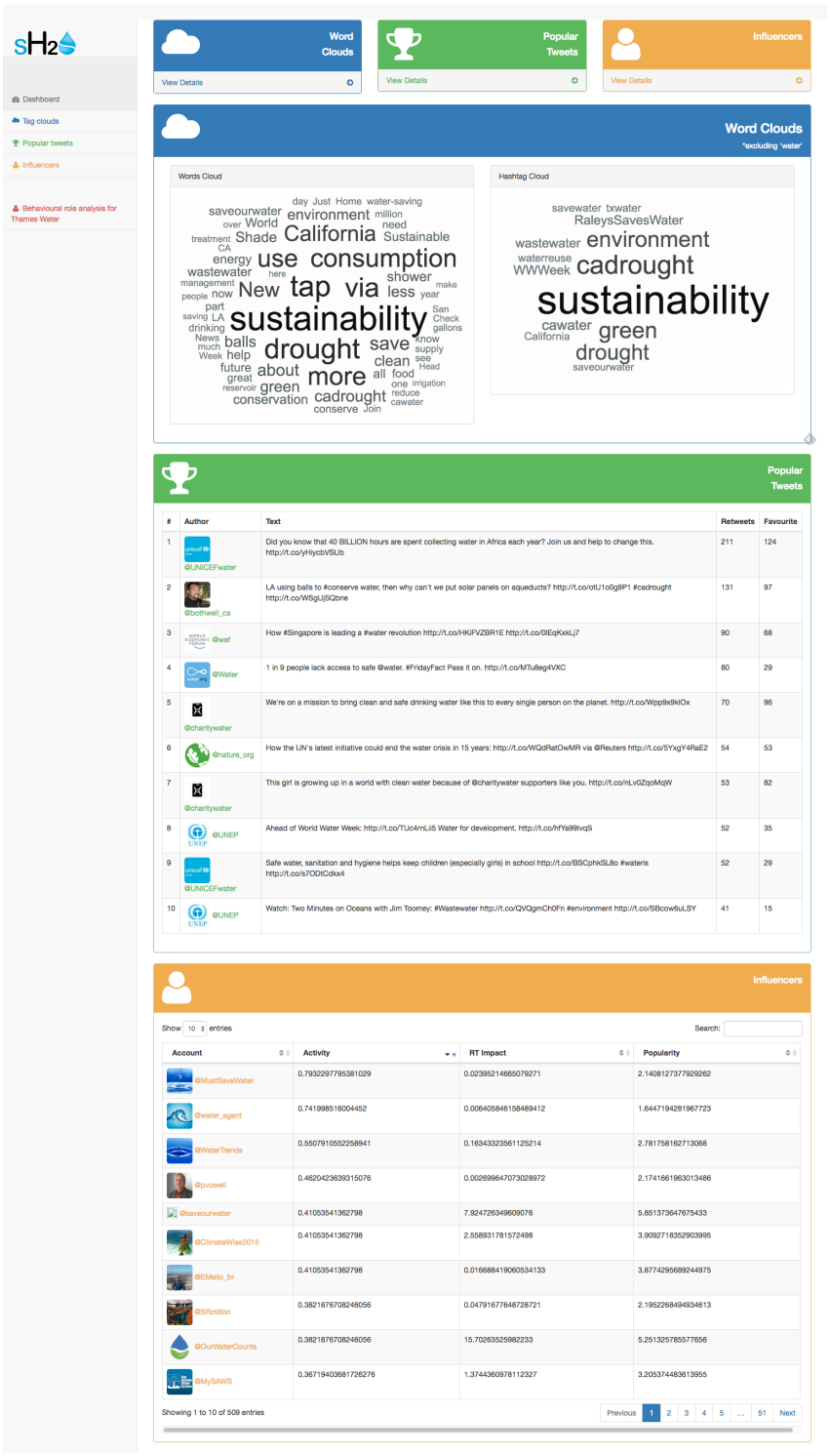


Figure 10. The content overview interface in the Social Dashboard.

5 Application of the SmartH2O method for behavioural role analysis to the Spanish case study

After the full implementation and integration of the developed method into the SmartH2O platform (the Social Network Crawler and Data Analyzer component) it has been applied to the Spanish case study (Valencia). This chapter describes the procedure and results of this application of the SmartH2O method for behavioural role analysis to the SmartH2O Spanish case study in Valencia. This application provides both an additional validation of the method validity and practical results informing the implementation of the SmartH2O Spanish case study. The first application to Thames Water Twitter community allowed for verifying the internal validity of the developed method. Its successful replication to the case study in Valencia (i.e. results containing comparable types of identified behavioural roles and cluster properties) allows us to test the stability of the behavioural patterns across the cases which verifies of the external validity of the methods; in particular, since this second case involves a completely different geographical, cultural and language community (a Spanish vs. English city/province).

5.1 Dataset construction

The dataset that was used for the first experiment was based on the followers of one single corporate water utility account. Whereas this experiment was ecologically valid in terms of its fit with the real-life application scenario this study envisions (e.g. the support of targeted multi-faceted communication campaigns), it is also limited, in the sense that communication behaviour could be the result of the users' specific relationship to the utility.

Using the social network crawler described in in Section 4, a dataset was constructed, based on 24 seed accounts. Seed accounts were manually selected, covering different private, and institutional (governmental, NGOs, cultural institutions, universities) accounts. This approach introduced variability into the dataset by including different relationships between private individual users and institutions. The complete list of seed accounts can be found in Appendix A.

For each seed account, first the list of all its followers and then all of their tweets were crawled for the timeframe between September 1st 2015 and August 31st 2016. To ensure geographical relevance to the SmartH2O Valencia case study, tweets were crawled only for accounts with users that were based in the metropolitan area of Valencia, as indicated in the location field in their user profile. In total, 6413 accounts were crawled. From this list, 4 accounts with excessively high activity were excluded (over 150 tweets per day; implying automated bots and spammers), as well as 194 accounts that were protected (e.g. hidden from non-friends and followers by the user). In the end, 6213 accounts could be used that yielded ca. 5.3 Mio tweets (5.341.151) for the analysis.

5.2 Procedure

The same method was applied to the crawled dataset as in the Thames Water study. The six behavioural dimensions were calculated using the same metrics as in the Thames Water study: activity, content contribution, dissemination, direct interaction, retweet impact, popularity, and attention triggering (see Section 3.2).

Spearman correlations were computed between the six behavioural dimensions (Section 3.2) to avoid non-normal distributions of the data to affect the results. Two sets of dimensions exceeded a .5 correlation coefficient: the correlation between direct interaction and attention trigger ($r(6213)=.571$) and the correlation between RT Impact and popularity ($r(6213)=.60$). Other correlations were significant, but relatively weak. All dimensions were introduced in the

cluster analysis. Since it couldn't be determined beforehand which of the two seemingly correlated dimensions should be kept and which one excluded, all were introduced in the cluster analysis. This would allow us to analyse the final results to better understand the nature of the correlation and if any of the two dimensions should be favourited or whether they, in spite of being correlated, still bring important aspects for differentiation (e.g. in analyzing the nature of the resulting clusters). Moreover, closely replicating the approach from first application in Thames Water case study ensured the comparability of the results.

An unsupervised clustering was performed on the user's behavioural dimensions in order to group users with similar behavioural patterns. Max-min normalization was used in order to give the same weight to each dimension. The clustering was performed by applying the K-means algorithm. The optimal number of clusters was iteratively determined by running the algorithm with a number of clusters ranging from k=2 to k=10, recording the value for the Davies-Bouldin Index. The results are shown in Figure 11.

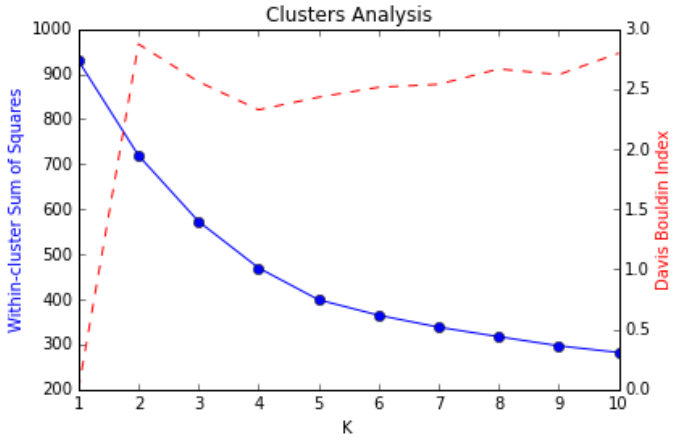


Figure 11. Determining the no. of clusters based on DB-index.

Comparison of the DB-indices suggests that the optimal number is k=4 (DB=2.36) or k=5 (DB=2.43). After qualitative inspection of the behavioural patterns, an optimal number of clusters was determined at k=5. Users were then clustered accordingly[2]. In the number of clusters is plotted against the DB-index and the within-cluster sum of squares (as a measure of homogeneity within the cluster).

5.3 Results

5.3.1 Clustering results

After clustering the users, the normalized averages for each of the dimensions were computed for each of the clusters (cluster centroids). The results are depicted in Figure 12 and Figure 13, showing the normalized averages, standard deviations, and within-group distribution for the different clusters. As can be seen in Figure 12 each cluster has a distinctive combination of the main dimensions. The within-cluster variability of individual values is comparable between the clusters, except for the "Authorities" cluster which in most dimensions tends to exhibit higher variability than other clusters (Figure 13).

[2] We actually tested both clustering for k=4 and k=5 and compared the difference to understand better the difference, The only major difference between clusters for k=4 and k=5 was in a cluster of passive users being included in k=5 and excluded in k=4. Since a cluster of rather passive users (observers or so-called lurkers) is typical for any online community, we have chosen the k=5 clusters as more correctly reflecting the structure of the community roles.

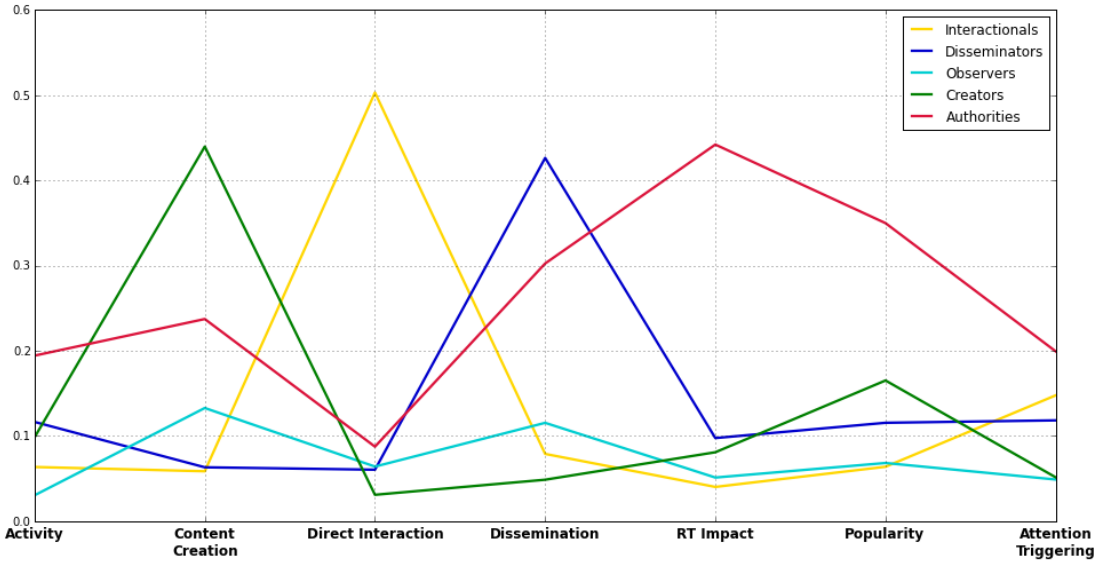
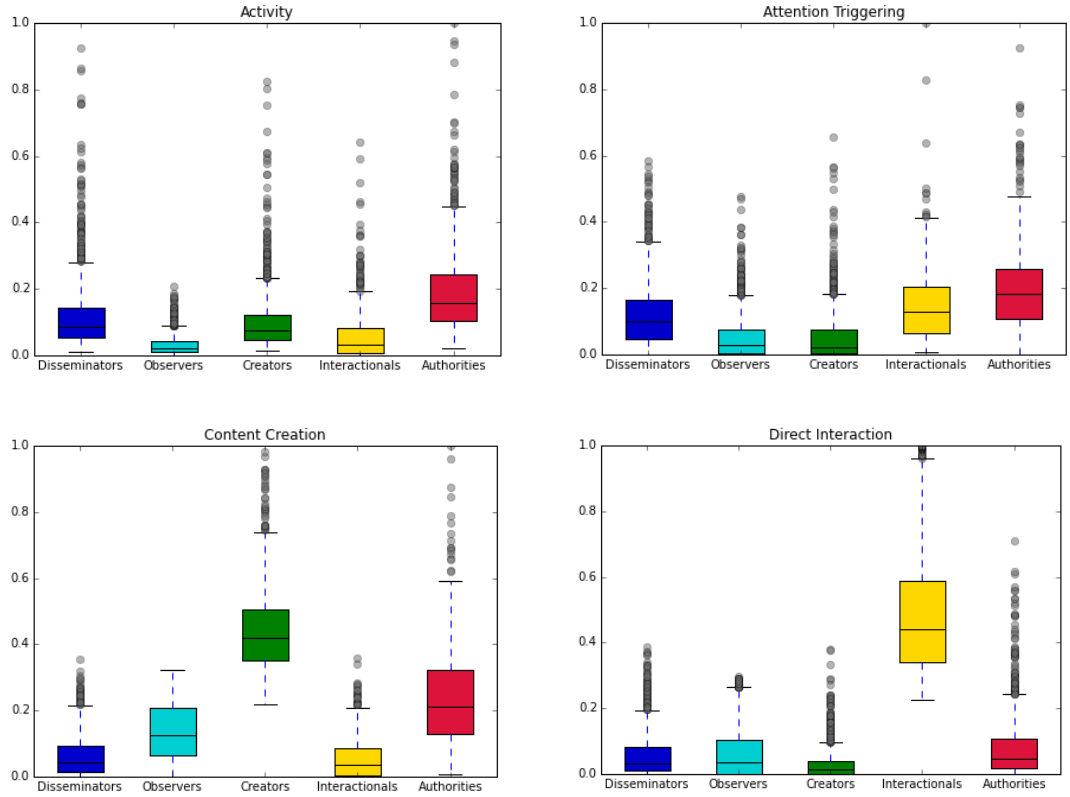


Figure 12. Average normalized values for behavioural dimensions per cluster.



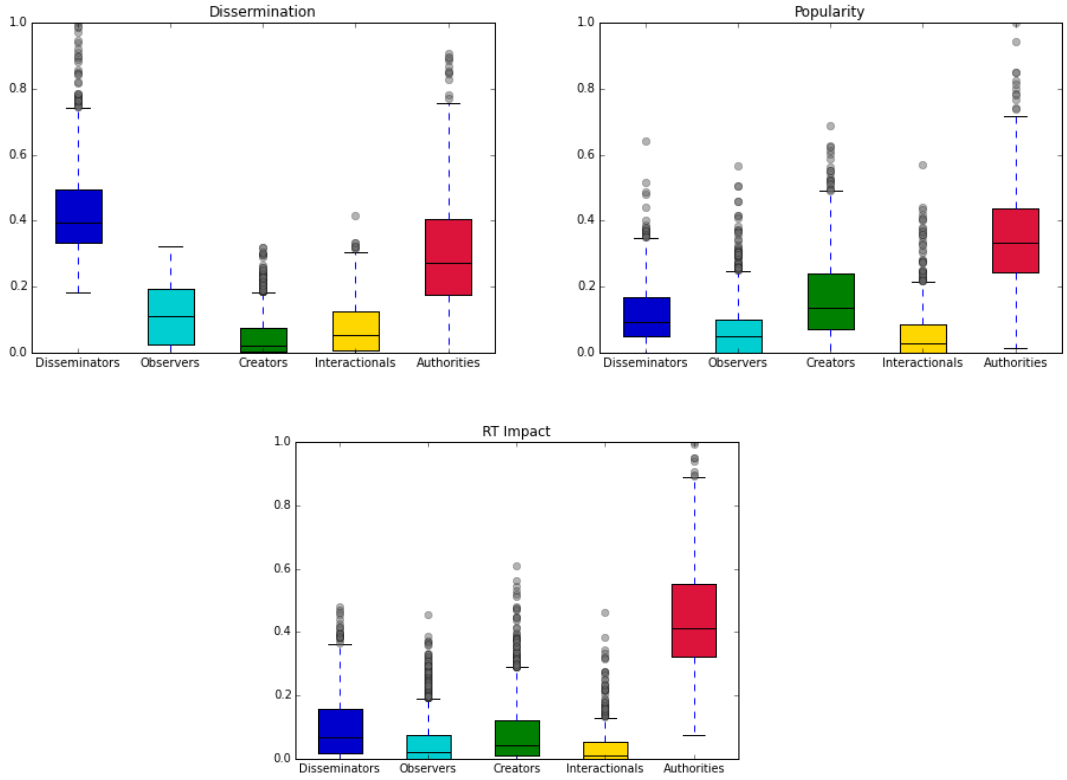


Figure 13. Distribution of values on behavioural dimensions across clusters.

Results reveal that the behavioural patterns across clusters is highly similar to the results we that were found for the Thames Water case dataset (Chapter 3). In fact, the clusters exhibit highly similar behavioural properties such that the same behavioural roles as in the first case study were found and assigned to the clusters. The identified behavioural roles corresponding to the clusters and the relative size of each cluster are displayed in Figure 14.

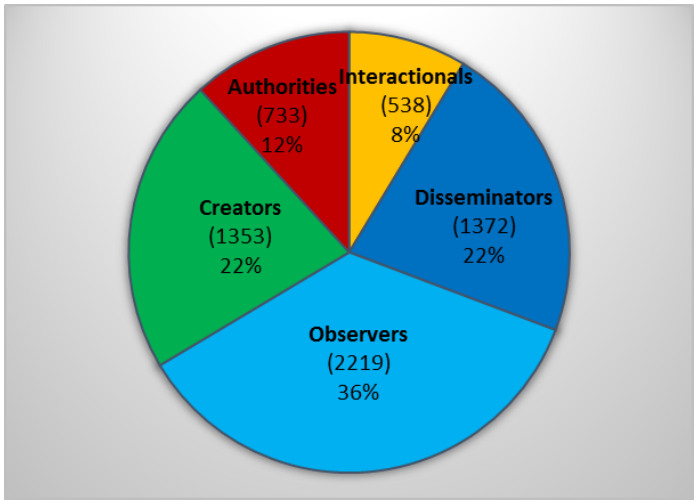


Figure 14. Number of users per cluster (Valencia).

5.3.2 Analysis and interpretation of behavioural roles

In this section we analyse the main properties of the individual clusters as well as the

structural commonalities and differences between them and the associated behavioural roles, with respect to their suitability to support a targeted spreading of communication campaigns.

Authorities: This cluster exhibits the highest activity, the highest RT impact and attention triggering. This makes them the ideal group for supporting the spread of communication campaigns as the content they create and spread frequently is very much retweeted. These users also exhibit a high value in dissemination, which further underscores the importance of this role for a wide and effective spread of messages noted by others. While very valuable, they comprise a small portion of users (12%), which suggests that a campaign cannot rely only on their activation, but needs to consider other appropriate behavioural roles to complement them.

Interactionals: users in this cluster exhibit a relatively low level of overall activity but the highest value of direct interaction among all clusters. This shows they frequently engage in conversations with others. They are second lowest on undirected dissemination of content to others (such as retweets). Their low RT impact and popularity but second highest attention triggering reflect their focus on small group communications and their frequent direct targeting of other users with @-mentions. This is important as such direct messages result in personal notifications being displayed by Twitter clients to the receiving users – which ensure that these messages are much more likely to actually be seen by the target recipients. In addition, such close-knit groups of users typically reflect strongly connected social communities, often relating to special interests; both of these increases the likeliness of engagement with content deemed of their interests (e.g. a SmartH2O dissemination campaign). Though accounting for the smallest portion of users (8%), these properties make them an important group to be addressed for further supporting the spread of targeted communication campaigns.

Disseminators: characteristic for this cluster are the highest dissemination and second highest activity and RT impact. This indicates that users in this cluster frequently propagate other users' content through the network. Their overall activity level is similar to the creators, but they score higher on popularity, and lower on attention triggering – suggesting they are less frequently targeting specific users in their tweets with @-mentions. Their widest reach for spreading the content of others alongside with high activity and RT impact make them potentially powerful amplifiers of the reach of communication campaigns, which should target and try to activate them in a focused way. They comprise a relatively large portion of users (22%), which additionally increases their potential contribution to supporting the spread of communication campaigns.

Creators: this cluster demonstrates relatively low activity but highest creation of new content. At the same time both their dissemination and direct interaction with their followers are very limited (the lowest values for all clusters). This suggests that they are unlikely to propagate messages created by others to their followers. While their RT impact and popularity are somewhat higher their attention triggering again exhibits the lowest value. This means that some of their own created content manages to create some interests (retweets) but their low activity, lowest dissemination (lack of re-tweeting the content of others) and low attention triggering make them unsuitable for the propagation of messages of others i.e. of a communication campaign.

Observers: here we observe low scores across all dimensions. This includes the lowest activity and very low overall dimensions except for a bit of own content creation and limited dissemination (becoming neglectable when compared to the very low overall activity). Accordingly, this cluster comprises the passive part of the Twitter user community, corresponding to 36% of users in this case study.

To support a more detailed analysis of the behavioural patterns that distinguish the clusters we also inspect the collected user statistics underlying the metrics in more detail. Results are displayed in Table 4. This allows us to analyse the corresponding behavioural patterns for the individual clusters in more details.

Table 4. User statistics for the Valencian case, divided by cluster.

Metric	Interactionals	Disseminators	Observers	Creators	Authorities
Tweets per day	1.5	3.1	.2	2.2	7.0
OT	114.6	152.1	38.4	65.5	917.2
CT	277.9	91.9	1.4	32.4	285.9
DT	119.3	846.2	35.1	92.1	1258.5
OT ratio	.2	.1	.5	.8	.4
CT ratio	.6	.1	.1	.0	.1
DT ratio	.2	.8	.4	.1	.5
Listed Count	9.0	18.9	7.7	31.1	123.8
Followers Count	313.0	518.3	311.0	985.1	5036.8
Friends Count	574.7	786.8	563.8	989.5	1856.8
Distinct Users Replied	411.1	332.3	33.1	305.3	396.6
Distinct Users Mentioned	105.2	76.3	15.6	62.9	279.7
Engagements (RT/FAV) Twuc abs	185.8	24.9	51.7	459.4	6671.1
Engagements (RT/FAV) per Twuc	1.1	2.0	1.4	.9	7.7
Twuc RT ratio	.2	.3	.2	.2	.7
Followers/(Followers + Friends)	.3	.4	.3	.4	.6
Engagements (RT) Twuc abs	58.7	98.9	17.4	207.4	3642.8
Engagements (RT) per Twuc	.3	.8	.4	.3	4.0

Authorities tweet most often (7 tweets per day), have the largest number of followers (5036) and friends (1856), and have the largest potential to reach their audience (observed from the ratio between followers and followers+friends, which is .6 for authorities). As such, authorities have the highest potential of the clusters to propagate messages and influence their audience. Interestingly, this is not only in line with the results for this cluster in the ThamesWater analysis but also the absolute values of some indicators are also very similar (e.g. 7 tweets/day here compared to 6,13 tweets/day in the ThamesWater analysis).

Interactionals send out tweets on average 1.5 times a day. Only 20% of these tweets is original content, while a much larger share (60%) are conversational tweets. Dissemination tweets only make up for 20% of their total number of tweets. Compared to the other clusters, their set of both followers (313) and friends (574) is relatively small, but the share of the users with whom the user interacts with is much higher than in the other clusters.

The **Disseminators'** most prominent feature is the highest DT ratio of all clusters, while still remaining moderately high scores on impact, popularity, and attention triggering. Their activity level is with 3.1 tweets per day second highest, after the Authorities. With a friends to followers+friends ratio of .4, the ability to reach their audience is second highest after the authorities, while being equal to the creators. The vast majority (80%) is made up of tweets that are produced by others and forwarded to his/her audience. Self-created tweets and conversational tweets only consume 10% of the total number of tweets each. Disseminators have the second largest share of their content retweeted in comparison to the other clusters, with only the Authorities receiving a higher score. This points out to the potential of disseminators for supporting the propagation of messages from communication campaigns.

Creators demonstrate a high OT ratio of 80%. Only 20% of the tweets are dissemination tweets or conversational tweets. With a little more than two tweets a day, they are more active than interactionals and observers, but less than disseminators and authorities.

Interestingly, even though they mostly create their own content, it yields few engagements from other users, as shown by the low engagements per Twuc rate (.3).

Observers create tweets only occasionally – on average once every 5 days (this is not only an order of magnitude lower than all other classes but also overall a low value for Twitter standards). If they do, 40% of their tweets is retweeting content of others (DT ratio), while 50% is original content. Observers rarely use Twitter for conversations with other users. The low followers to friends+followers ratio is testimony to the limited potential of this cluster to reach and engage their audience.

6 Lessons learnt from comparative analysis of two application cases

In the previous Sections we have described the application of the SmartH2O method for behavioural role analysis in Twitter communities in two field cases. Especially the successful replication of the method application to the Spanish case study underscores the validity of our approach and of the developed methods: this case study is very different by essential properties, such as the geographical community, cultural background and language, yet yielded the same behavioural roles and their key properties as in the first case. The analysis of the results also shows how the identified behavioural roles and associated behavioural patterns can be used to support targeted spreading of communication campaigns, by taking into account the behavioural roles users display on social networks. In this section we comparatively analyse the two cases and draw preliminary lessons for the set-up of such communication campaigns.

6.1 Comparative analysis

The results have revealed highly similar clustering of users and associated behavioural patterns, as can be seen from the figures below. The cases primarily differ in the amplitude of the metrics, but the pattern remains the same across the cases (with some similarity even in the amplitudes, e.g. for the Authorities clusters).

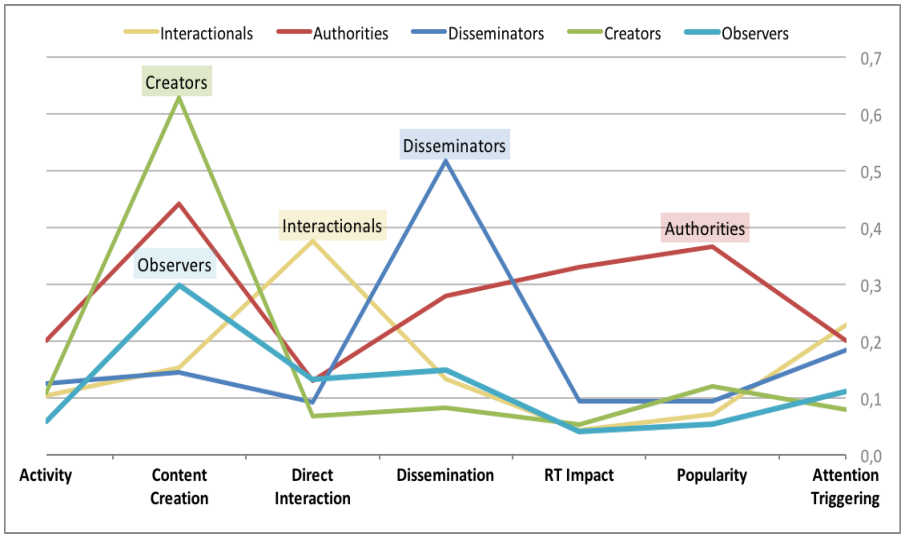


Figure 15. Values of behavioural dimensions of cluster centroids (Thames Water).

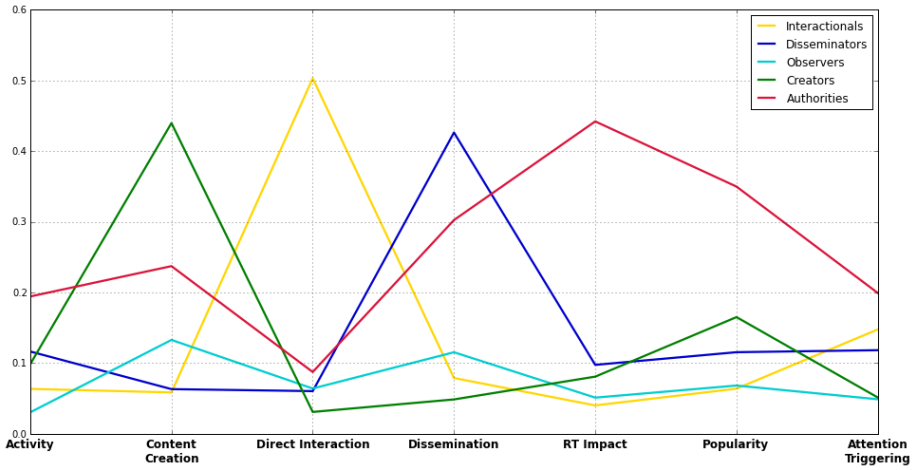


Figure 16. Values of behavioural dimensions of cluster centroids (Valencia).

The relative size of the clusters also displays a similar pattern, with some (positive) differences with respect to the size of active clusters, as can be seen from Table 5.

Table 5. Comparison of cluster sizes across cases.

Cluster	Thames water	Valencia
Authorities	9%	12%
Disseminators	16%	22%
Creators	21%	22%
Interactionals	19%	8%
Observers	35%	36%

The comparison demonstrates that the share of observers, creators, and authorities is more or less the same across the two cases. Overall, the proportion of Observers as passive users is almost the same in both cases (35% vs. 36%). This suggests that the overall group of active users is the same in both contexts. The cases differ in the internal distribution of the types of active users among themselves The biggest difference can be observed in the much larger portion of Interactionals in Thames Water than in the Valencia case (.8% vs. 19%). On the other hand, the portion of Disseminators is somewhat larger in the Valencia than in the Thames Water case (22% vs. 16%). This suggests that in the Thames Water case there is a much larger portion of users who engage in closed-group interactions (twice as many), while in the Valencia community, the behavioural roles exhibiting broad, un-directed communication patterns are more pronounced (6% more Disseminators and 3% more Authorities than in the Thames Water case)[3].

To further inspect the similarities and differences between the two cases, we have put together the user statistics most important for communication campaign planning side by side, as shown in Table 6.

[3] One can only speculate about possible reasons for such differences; investigating that would require a separate sociological study in its own right, which could be interesting but is way out of scope of this project (and is not related to the goals being pursued).

Table 6. Comparing selected user statistics between application cases.

Metric	Thames Water					Valencia				
	Auth.	Dissem.	Crea	Interact	Obs	Auth.	Dissem.	Crea	Interact	Obs
Tweets per day	6.1	2.9	2.2	2.1	.7	7.0	3.1	2.2	1.5	.2
OT ratio	.5	.2	.8	.2	.5	.4	.1	.8	.2	.5
CT ratio	.2	.2	.1	.6	.3	.1	.1	.0	.6	.1
DT ratio	.3	.6	.1	.2	.3	.5	.8	.1	.2	.4
Twuc RT ratio	.5	.2	.1	.1	.1	.7	.3	.2	.2	.2
Followers/(Followers + Friends)	.7	.3	.4	.3	.3	.6	.4	.4	.3	.3
Engagements (RT/Fav) per T_{wuc}	4.2	.9	0.4	.6	.4	7.7	2.0	.9	1.1	1.4

The table demonstrates strongly similar behavioural patterns across cases, which is in line with the overview of the (normalized) key metrics presented at the beginning of this section. Most differences on the indicated metrics are within a .1 range. This is an interesting finding as it suggests that the identified behavioural roles, and metrics to measure them, have behaved robustly across two very different cases.

For the disseminators, the DT ratio is the most prominent measure. Average values for Valencia are even more pronounced (.8) then in the Thames Water case (.6.). A similar pattern can be observed from ratio of tweets with original user content that gets retweeted, which is a distinguishing measure for Authorities. At .7 in Valencia, this value is .2 higher than in Thames Water (.5). This suggests that, combined with the high followers/ (followers+friends) and higher no. of engagements per Twuc ratio. Authorities have a strong potential to influence their audience and attract attention to the content they are posting. In addition, the non-normalized activity values are also very similar for the Authorities (7 tweets/day in Valencia vs. 6,13 tweets/day for Thames Water), which suggests that this type of users also share structurally similar activity pattern.

6.2 Lessons learnt and recommendations

The presented results suggest that the proposed method for analysis of behavioural roles in Twitter can identify a set of user types (and users) that can support information dissemination in Twitter campaigns in a multi-faceted, targeted manner. The experimental application to two datasets of 6,3 million and 5,0 million tweets (respectively) in two application cases has uncovered a robust set of user groups exhibiting behavioural roles suitable for this purpose. The identified roles are similar to those initially hypothesized based on a critical synthesis of existing literature, but are more closely related to the specific needs of the stated application context of setting up multifaceted communication campaigns. Having applied the method to two cases covering customer bases of two different water utilities also ensures direct applicability of the results for the campaigns in the water utility sector. The identified roles, associated user groups on Twitter and the properties of their behavioural patterns can readily be used to design strategies for effectively spreading Twitter-based communication campaigns (e.g. for promoting water saving) in a targeted manner.

The consistent findings between the cases suggest that the introduced behavioural dimensions and metrics are suitable for capturing behavioural patterns that can support targeted information dissemination. The identified roles and their distribution in the Thames Water case and the same roles and their distribution found in the Valencian case demonstrate that much larger portion of active users, suitable for supporting Twitter-based communication campaigns, can be identified with this method in comparison to usual

influencer approaches. Accordingly, the described findings provide direct support for designing targeted communication campaigns by water utilities within the SmartH2O context.

Furthermore, the technical implementation of the developed method and its integration into the SmartH2O platform in form of the component SmartH2O Social Network Crawler and Data Analyzer provides an interactive tool that can be readily use to readily repeat or perform the community behavioural roles analysis on different cases. The interactive visual interface allows intuitive examination of the main metrics of the individual behavioural clusters as well as the browsing and identification of most relevant individual users. In this way, the obtained results and lessons learned have been provided in a tool that is easily applicable to different water saving and awareness campaigns.

Recommendations

Based on the results presented in this deliverable and the results we have obtained from the assessment of the SmartH2O incentive model, several recommendations can be defined for promoting water saving through behavioural change apps and social network communication campaigns. We first present recommendations related to behavioural role-based communication campaigns, following from the approach presented in this deliverable. We then provide recommendations on the most effective leverage to stimulate water saving, based on insights gained from user feedback on and users' interaction with the SmartH2O application and its incentive model (see D4.4).

Behavioural role-based communication campaigns can support app-based incentive models to promote water saving in larger communities

The role-based approach presented in this deliverable can support water saving programs that combine app-based incentives with communication campaigns that seek to incentivize users to sign up for these water saving awareness applications. When the initiator of the water saving campaign (e.g. a water utility, or a city) already has a Twitter account with a follower base, different behavioral role types can be addressed to leverage different types of influence. For example, authorities can be employed to propagate recruitment messages quickly to a large number of users, while interactionals are useful to reach special interest groups with strong connections between the users. These interactionals can then support the recruitment of users within this group. This is particularly effective when the 'interactional' is already a user of SmartH2O.

If such a follower base is not available, Twitter advertisements can be used once seed accounts of influential private users or local interest groups have been identified, and subsequently a relevant dataset of followers has been crawled and clustered. With this dataset as a basis, an advertisement campaign can be set up, whose targeting is based on behavioral roles that allow to effectively target the most suitable users.

Extend the impact of water saving tips and (aggregated) feedback beyond the users of water conservation apps

To increase the impact of water saving apps, the developed behavioural roles analysis can be used to support communication campaigns that target households that couldn't be persuaded to sign-up for using such applications. For this share of the audience, different incentives for water saving need to be sought. In that case, communication campaigns can be used to directly increase water saving awareness through the dissemination of e.g. water saving tips, or aggregated consumption feedback at the level of the community as a whole. To increase the reach of such campaigns, the behavioral roles can be employed. For example, disseminators and authorities can be used to disseminate tips within the community at large, while 'creators' that are already using a water conservation app can be addressed to create messages with tips from these applications.

In addition to recommendations related to the SmartH2O approach for behavioral role analysis on social networks, the interim evaluation of the incentive model (reported in D4.4) yields insight into the most effective leverage to stimulate water saving behavior. With respect

to the design of water saving applications, the following recommendations can be formulated.

Employ a holistic incentive model with different types of incentives to address different types of users (e.g. pragmatic and hedonic, competitive and achievers).

Both the requirements analysis, and the analysis of user feedback and user activity highlight the importance of an incentive model that is capable of stimulating different types of users. Users with primarily pragmatic expectations can be motivated by for example detailed consumption feedback, and tips, whereas users for whom fun-of-use is important can best be motivated by hedonic, gamified elements. The SmartH2O incentive model was designed to appeal to both types of users by integrating the different elements that cater to the specific preferences of these different types of users. The consistently positive user feedback to both the more pragmatic features, and the gamified elements, as well as the observed usage patterns provide evidence for the suitability of such a mixed pragmatic-hedonic approach. This is also supported by the results from the user activity analysis, where different usage patterns were observed, especially relating to the use of the competitive elements. For example, while some users were highly active and wanted to remain on top of the leaderboard, others users reported being less motivated by such competition, but more by achievements (e.g. saving water, receiving points for activity on the portal).

Design attention triggering mechanisms to keep active users engaged and to re-engage passive users.

The level of activity on the SmartH2O portal succeeded in exceeding the common frequency of interaction between utilities and their customers. However, the results show that this could be even further improved by pro-actively triggering the attention of the users. These mechanisms can serve to reactivate passive users, inviting them back to the application. Alternatively, for currently active users, motivating achievement summaries can be sent that incentivize them to remain active. For users whose consumption is increasing, notifications can be used that draw the attention to the increased consumption and encourage the user to take action. In particular, moible apps – such as the one developed in SmartH2O (described in D4.4) – are effective vehicles to send notifications to incentivize users to return to the application, to remain continuously active in water saving, or to start saving water again.

Design consumption visualizations to appeal to users with different motivations

In research on water consumption behavior and in the requirements and evaluation analysis in SmartH2O it was found that users differ in terms of their level of interest in their water consumption and the goals they pursue with using water consumption applications. In particular, the requirements analysis in SmartH2O has shown that water consumption feedback should include different motivational affordances to appeal to at least two types of users:

- **Data affine users**, who are really interested in consumption details. Consumption visualizations for these users should enable the interactive inspection of consumption information. Changing timespans and levels of detail, comparison against historic values, and comparison against similar households are examples of features that appeal to these users.
- **Users for whom fun-of-use is important**. These users are less interested in the precise details, but are more motivated by a visually appealing “at a glance” visualization that fosters playful engagement. The use of visual water-related metaphors allows for a semantic mapping of abstract consumption figures to concepts that fit within the frame of reference of the user is particularly effective to easily visualize consumption as well as to demonstrate the impact of water consumption. Playful engagement can be encouraged with gamified competition mechanisms, social comparison against other users, and self-set consumption goals to give the user a sense of achievement.

In SmartH2O, this distinction has been implemented by offering an interactive bar chart for the data-affine users, and a consumption visualization based on a pipe metaphor for users for whom fun-of-use is important. This was combined with a display of a number of swimming pools, to help users understand how much water they are using on a yearly basis. Self-reported user motivation and an analysis of activity levels has demonstrated the positive uptake of both visualizations, providing evidence for the effectiveness of the motivational affordances embedded in the design.

Effective functionalities for aplications supporting behavioural change for water saving

Finally, the results of the incentive model evaluation suggest that the following functionalities are well-suited to motivate users to save water and should thus be considered when designing applications for water saving through behavioural change:

- **Consumption visualizations.** Visualizations can either be **metaphor-based**, or can contain traditional **bar charts**. The former was intended for easy understanding and to make the impact of water consumption tangible, whereas the latter allows for detailed inspection of consumption levels.
- **Leaderboard in combination with physical rewards** has shown to create a real competition between users, while in the process these users are exposed to incentives to save water.
- **Self-set water consumption goals**, to create commitment towards water consumption reduction
- **Water saving tips**, to provide users with the necessary knowledge and skills to save water, and to stimulate self-confidence to do so.

7 References

[Beguerisse-Díaz et al., 2014] Beguerisse-Díaz, M., Garduño-Hernández, G., Vangelov, B., Yaliraki, S. N., and Barahona, M. 2014. Interest communities and flow roles in directed networks: the Twitter network of the UK riots. *Journal of The Royal Society Interface*, 11, (Dec 2014), 101.

[Brandtzaeg & Heim, 2011] Brandtzaeg, P. B., and Heim, J. 2011. A typology of social networking sites users. *International Journal of Web Based Communities*, 7, 1 (Jan 2011), 28-51.

[Cha et al., 2010] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. 2010. Measuring user influence in Twitter: the million follower fallacy. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (Washington, D.C., May 23 – 26, 2010), Menlo Park, CA: AAAI Press,10-17.

[Dugué & Perez, 2014] Dugué, N., and Perez, A. 2014. Social capitalists on Twitter: detection, evolution and behavioral analysis. *Social Network Analysis and Mining*, 4, 1 (Mar 2014), 1-15.

[Enge, 2014] Enge, E. 2014. Twitter Engagement Unmasked: A Study of More than 4M Tweets. Retrieved Feb 9, 2014, from: https://www.stonetemple.com/twitter-engagement-umasked/

[Fagnan et al., 2014] Fagnan, J., Rabbany, R., Takaffoli, M., Verbeek, E., and Zaïane, O. R. 2014. Community Dynamics: Event and Role Analysis in Social Network Analysis. Advanced Data Mining and Applications, 85-97. Springer International Publishing.

[Ghosh et al., 2012] Ghosh, S., Viswanath, B., Kooti, F., Sharma, N. K., Korlam, G., Benevenuto, F., and Gummadi, K. P. 2012. Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st international conference on World Wide Web* (Apr 2012), ACM, 61-70.

[Gleave et al., 2009] Gleave, E., Welser, H. T., Lento, T. M., & Smith, M. A. 2009. A conceptual and operational definition of social role'in online community. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference* (Big Island, HI, January 5 – 8, 2009), IEEE, 1-11.

[Golder & Donath, 2004] Golder, S. A., and Donath, J. 2004. Social roles in electronic communities. *Internet Research,* 5, 19-22.

[Guimerà & Amarral, 2005] Guimerà, R., and Nunes Amaral, L. A. Functional cartography of complex metabolic networks. Nature 433 (2005), 895–900.

[Honeycutt & Herring, 2009] Honeycutt, C., and Herring, S. Beyond microblogging: Conversation and collaboration via twitter. In System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on (Jan 2009), 1–10.

[Maulana & Tjen, 2013] Maulana, A. E., and Tjen, S. 2013. From the angels to the screamers: User segmentation and e-WOM in social media. In International Proceedings of Economics Development and Research, 55, 67-71.

[Pal & Counts, 2011] Pal, A., and Counts, S. 2011. Identifying topical authorities in microblogs. In *Proc. of 4th ACM International Conf. on Web search and data mining*. (Hong Kong, China,), 45-54.

[Rowe & Alani, 2012] Rowe, M., and Alani, H. 2012. What makes communities tick? Community health analysis using role compositions. In *Proc. of PASSAT/SocialCom 2012,* (Sep 2012), 267-276. IEEE.

[Scripps et al., 2007] Scripps, J., Tan, P.-N., and Esfahanian, A.-H. 2007. Node roles and community structure in networks. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, (San Jose, CA, USA, August 12 - 15, 2007) WebKDD/SNA-KDD, 26–35.

[Suh et al., 2010] Suh, B., Hong, L., Pirolli, P., and Chi, E. H. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, 177-184. DOI = http://doi.acm.org/10.1109/SocialCom.2010.33.

[Tinati et al., 2012] Tinati, R., Carr, L., Hall, W., and Bentwood, J. 2012. Identifying communicator roles in twitter. In *Proceedings of the 21st international conference companion on World Wide Web* (Apr 2012) ACM, 1161-1168.

[Tyshchuk et al., 2013] Tyshchuk, Y., Li, H., Ji, H., and Wallace, W. 2013. Evolution of communities on Twitter and the role of their leaders during emergencies. In *Proceedings of 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*, 727–733.

[Wang et al., 2012] Wang, X., Liu, H., Zhang, P., and Li, B. 2012. Identifying information spreaders in twitter follower networks. *School of Computing, Informatics, and Decision Systems Engineering*. Arizona State University, Tempe, AZ.

[Welch et al., 2011] Welch, M. J., Schonfeld, U., He, D., and Cho, J. 2011. Topical semantics of twitter links. In *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11,* 327.

[Weng et al., 2010] Weng, J., Lim, E.-P., Jiang, J., and He, Q. Twitterrank: Finding topic-sensitive influential twitterers. In Proc.of 3rd ACM International Conf. on Web Search and Data Mining - WSDM '10, (New York, NY, USA, 2010), ACM, 261–270.

Appendix A

List with seed accounts that together with their followers have been used for the crawling of the Valencian dataset:

Username	name	Tweets	I follow	Followers	account type
100x100valencia	Valencianos	25000	74500	83200	organisation
UV_EG	Universitat València	6738	400	46300	organisation
mediambientcat	Medi Ambient (Catalonia)	27000	8485	18100	organisation
metrovalencia	Metrovalencia	9693	3738	14700	organisation
Bioparc	Bioparc Valencia	14000	2368	12700	organisation
Podem_	Twitter oficial de Podem a la Comunitat Valenciana	13400	4769	11400	organisation
valencia	Costa Valencia	2.098	4.728	5.014	organisation
informavalencia	Magazine digital informa Valencia	16400	1013	4338	organisation
valenciacity_es	Valencia City	3340	643	4029	organisation
AndreuEscriva	Andreu Escrivà	107000	514	3542	person
GVAParcs	Parcs Naturals de la Comunitat	3480	248	3403	organisation
Bio_Valencia	BioValencia	11200	1422	3254	organisation
mariajpico	Maria Josep Picó	4854	1601	2395	person
samarucdigital	samarucdigital	4165	1559	2174	organisation
GVAagroambient	Conselleria d'Agricultura, Medi Ambient, Canvi Climàtic i Desenvolupament Rural	1785	291	1668	organisation
valencia_lab	ValenciaLAB	2363	836	1649	organisation
Economia_3	Economía3	6529	211	1617	organisation
Elenetcc	Elena Cebrian Calvo	1214	773	1574	person
JardiBotanic_UV	Jardí Botànic UV	3369	645	1539	organisation
AVAESEN	Asociación Valenciana de Empresas del Sector de la Energía	5984	646	1217	organisation
AMAUPV	Medio Ambiente UPV	1023	160	1155	organisation
rosadg_	Rosa Dominguez Gomez	7057	790	1019	person
Oceanografic_vl	Oceanogràfic VLC	507	71	593	organisation
AVACUconsumo	Asociación Valenciana de Consumidores y Usuarios	924	155	506	organisation