Received: 21 December 2016 | Revised: 24 March 2017 | Accepted: 29 June 2017

DOI: 10.1111/caim.12228

SPECIAL ISSUE ARTICLE

Digital anthropology as method for lead user identification from unstructured big data

Vella Verónica Somoza Sánchez[1] | Davide Giacalone[2] | René Chester Goduscheit[1]

[1] Department of Marketing & Management, University of Southern Denmark, Odense, Denmark

[2] Department of Technology and Innovation, University of Southern Denmark, Odense, Denmark

Correspondence
Vella Verónica Somoza Sánchez, Department of Marketing & Management, University of Southern Denmark, Campusvej 55, 5230 Odense, Denmark.
Email: vella@sam.sdu.dk

This paper focuses on lead user identification in an open data context through the principles of digital anthropology (DA). The approach is demonstrated using a case study from the entertainment industry: the festival of Tomorrowland. Following the principles of DA, the specific data structure of the case selected is studied. This structure follows the tribal roles of Cova and Cova, which are then compared with the lead user characteristics of Belz and Baumbach. The different characteristics and roles are used afterwards to establish the user types that may be found in the big data set of the case study. Unlike earlier methods, such as netnography, which focuses on a specific user type as the main source of insights, our results show that DA requires a change of focus from specific individuals to groups of individuals with the capacity to produce the insights of a lead user. DA shows how all the roles in the data structure present different possibilities for innovation. The application of DA to big data enables the extension of analysis to whole populations instead of specific samples, bringing about new possibilities to model and extract the insights of consumers publicly exposed in the digital media.

1 | INTRODUCTION

The identification of lead users is a relevant topic due to their potential as innovation sources (Belz & Baumbach, 2010; von Hippel, 1986). The importance of lead users for companies resides in the insights they bring to the development of innovation processes, as well as the detection and prevention of problems related with products (Stockstrom, Goduscheit, Lüthje, & Jørgensen, 2016). In particular, previous work has focused on lead users as sources of knowledge regarding product use, customer needs, and demand trends (Chatterji & Fabrizio, 2014; Cohen, Nelson, & Walsh, 2002; Laursen & Salter, 2006; Stockstrom et al., 2016; von Hippel, 1988). Thus, different methods to identify the right user that may help to tap into this knowledge have been proposed in recent years (Lüthje, 2004). Traditional methods for lead user identification include three main techniques: pyramiding, screening, and broadcasting (Stockstrom et al., 2016). These procedures have in common the aim of identifying the most relevant user for a given product or technology that will experience the same problems that the market will face before they occur, having at the same time the potential to find solutions beforehand (Belz & Baumbach, 2010; Stockstrom et al., 2016). These methods present some drawbacks, such as high cost (especially screening due to the need to contact the whole study population) and excessive reliance on self-assessment. The current research presents an alternative method that takes advantage of the data-rich environment available in the digital universe in which companies and individuals are embedded (Bharadwaj & Noble, 2015).

During the last 30 years, the access to information and data has changed to a level in which its extent, management, and continuous growth have generated a new interest in so-called “big data” (Feldman, Kenney, & Lissoni, 2015; Mayer-Schönberger & Cukier, 2013). Although the exploitation of big data for lead user identification has been proposed, actual tools enabling its efficient manipulation and exploitation are lacking (Feldman et al., 2015). Situated within this context, the current article presents an alternative method coming from anthropology. Digital anthropology (DA) (Miller & Horst, 2012) is a sub-discipline of anthropology, proposed theoretically as a method by Boellstorff (2012). The potential of DA for the study of big data resides in the fact that its principles establish the need to study in a holistic manner the network of relationships among the data. In this sense, the authors invite its further use and development for other applications. In the present paper, DA is adapted and empirically applied for the study of lead user identification. In this context, our proposed method keeps the positive characteristics of traditional methods (e.g., broadcasting and screening) while addressing some of their drawbacks. At the same time, we address some of the disadvantages of prior methods for lead user identification in rich-data environments (e.g., netnography), by proposing an approach better suited for an open data context. The principles of DA state the importance of taking into consideration one case at a time and to study its internal structure in order to establish results for analysis (Miller & Horst, 2012). Therefore, in this paper we focus on an empirical case study to discuss the application of DA and to evaluate its suitability for lead user identification from open data. Hence, the main objective of this research is to complement the literature on lead user identification towards new possible methods, such as DA, that benefit from unstructured big data; more specifically in this case, "from the rich data environment that characterizes social media".

The results of the current research have theoretical and managerial implications. From a theoretical perspective, DA contrasts with other methods such as netnography by improving its efficiency through taking into account the relative importance of all the user roles for lead user identification. In this sense, not only devotees are considered as the insight producers for the community, but all user clusters gain interest. Additionally, the study shows how the social media architecture impacts the content that the users may produce and therefore attracts different user types. In this sense, different channels are suitable to produce different lead user characteristics, attracting thus users presenting those specific characteristics. From a managerial perspective, DA is an improved method for lead user identification that allows not just the identification of the right user, but already the capitalization of the ideas that these users publicly expose on social media. Another advantage of this approach of lead user identification entails the avoidance of over-reliance on user self-assessments (advantageous over screening and pyramiding).

2 | LITERATURE REVIEW

2.1 | Conventional methods for lead user identification

There are many different methods for lead user identification. Traditional methods include pyramiding, screening, and broadcasting, among others. Pyramiding is a sequential search process. It starts by asking an initial contact to provide references to one or more persons whom the initial contact considers to exhibit a higher level of the attribute or quality being sought or to have better information regarding who such people might be (von Hippel, Franke, & Prügl, 2009; von Hippel, Thomke, & Sonnack, 1999). During the next step, these persons are then approached and asked the same questions. This is repeated until individuals with sufficiently high levels of the special quality have been identified. As such, pyramiding may prove useful when trying to identify individuals with high levels of a special quality or attribute. Due to its reliance on references, pyramiding is primarily affected by what people know about others (von Hippel et al., 2009), and therefore relies on the assumption that respondents can correctly evaluate other people′s expertise, which may or may not hold true in reality (Brem & Bilgram, 2015).

Screening and broadcasting, on the other hand, are parallel search processes. In screening, the searching party administers a questionnaire or conducts interviews with all the members of the population in order to collect information about their characteristics. This implies, first, that the population is defined as well as its boundaries, and second, that data is collected (e.g. through a survey) from every member of this user population to determine whether they possess the desired characteristics and knowledge. If a firm manages to contact the entire population and to receive answers from all subjects, screening is an effective approach. Using this method, a company will definitely find the users who display the pre-defined characteristics being sought. One pitfall is that these two preconditions are usually not met (Poetz & Prügl, 2010). In addition, the efficiency of this approach may be low, especially if the cost of contacting members from the user population is high, if they are very hard to reach, or if the characteristics being sought are increasingly rare in the population. In these situations, a traditional screening process can be difficult and expensive (Prügl & Schreier, 2006; Sudman, 1985; von Hippel et al., 2009).

In contrast to screening, broadcasting is carried out by "disclosing the details of the problem at hand and inviting the participation of anyone who deems themselves qualified to solve the problem" (Jeppesen & Lakhani, 2010, p. 1016). That is, experts self-select to reveal their expertise after having been made aware of the problem/topic. Broadcasting therefore does not require the searching party to define the boundaries of the target population or to obtain data from every individual in it. With regard to the information they seek, screening and broadcasting both focus on what people know about themselves, as they fundamentally ask "do you have the special quality we are looking for?" (von Hippel et al., 2009). Although broadcasting is slightly more efficient than screening, this heavy reliance on self-assessment may be a problem as people are known to often overestimate their abilities across different domains (e.g., Kruger & Dunning, 1999).

In recent years, new alternative methods based on open data have been proposed to address some of the drawbacks of the traditional methods. For example, the use of netnography has been presented as a way to identify lead users from emerging sources of data such as social media (Belz & Baumbach, 2010). Netnography is a hybrid term made up of internet and ethnography. It is a new method to scrutinize an online community data (Kozinets, 2010). Netnography takes its point of departure in research methods like observation as a means to study cultures and communities and apply it to the various activities that are carried out online. Netnography uses publicly available information in online communities to analyze and understand consumer needs, consumer trends, and consumer behavior and its influences (Kozinets, 2002). The substantial growth of the amount of data in online communities has made these platforms a useful means to search for lead users. Previous studies have shown that lead users are actively contributing to the generation of knowledge about existing products, and have sought to employ online communities as a way to formulate needs and preferences in relation to the products or services (Sawhney, Verona, & Prandelli, 2005). In this sense, netnography has been suggested as a valuable method to gain insight into the online communities and to potentially identify lead users.

2.2 | Lead users′ characteristics

Belz and Baumbach (2010) and Lüthje (2004) present six different characteristics that define a lead user. These characteristics are related to specificities or needs of the users in relation to a target product. The first characteristic regards new needs that are yet to be addressed by the current market, defined as "*ahead of trend*". The second characteristics is the "*dissatisfaction*" of the users regarding the current product in relation to its attributes and perceived performance, which can be used to develop and improve current products (von Hippel, 1988). The third characteristic concerns the "*use experience*" of interacting with a product that emerges through use and which helps to identify its problems (Lüthje, 2004). The fourth characteristic is "*product knowledge*". This level concerns deeper user knowledge about the product or service and its architecture in relation with the industry (Lüthje, 2004). The fifth characteristic considers the "*involvement*" and commitment of the user with the specific market. Finally, the last characteristic evaluates the perceived "*opinion leadership*" of users regarding the service or product object of study (Kratzer & Lettl, 2009).

3 | DIGITAL ANTHROPOLOGY FOR LEAD USER IDENTIFICATION

The literature shows several examples where data from online communities has been analyzed using netnography (Belz & Baumbach, 2010; Kozinets, 2002, 2010). However, netnography typically focuses on traditional online communities from the Web 1.0 era, i.e., online communities set around a blog or forum where the users discuss a given theme or product (Murugesan, 2009). The appearance of the Web 2.0, and the social media online communities along with it, has significantly changed the context of these relationships, as well as the interactions among the community members. Accordingly, the use of netnography may present difficulties when directly applied to these communities. Some of these difficulties concern the sampling of the communities: the popularity of social media implies the presence of data from a much larger amount of users, making netnography inefficient regarding time cost. Seeking to find an approach that overcomes the drawbacks of the current methods for lead user identification, as well as the inclusion of big data analysis due to the characteristics of the data produced in social media, the current study establishes the grounds for using DA as a method.

The case study selected from the social media YouTube presents such a quantity and velocity of production of comments that they may be regarded as big data (McAfee & Brynjolfsson, 2012). Such contexts present a series of difficulties for the current approaches. By contrast, DA allows a better accountability of the data through triangulation, strengthening the validity of the research results (Bosch-Sijtsema & Bosch, 2015).

The principles of DA have been formulated with specific focus on the online media (Miller & Horst, 2012). DA establishes that different sources of data may highlight different aspects (and characteristics) of a user, and thus, each source needs to be studied attending to its structure to determine which type of results it can produce. For example, the quantity of text that may be produced per comment in social media such as Facebook is far larger than what may be found in others like YouTube. This data should be studied using triangulation aiming towards a holistic understanding of the network of relationships among the data. This is particularly relevant in the study of big data, where the study of the whole population seeks to represent the big picture following the interconnections created along the network of data. Additionally, DA states that digital culture is produced locally; the type of behaviors and interactions produced in a certain website do not have to be reproduced equally on the Internet. This is in stark contrast to netnography, which proposes a homogenizing series of steps for analyzing online communities.

The anthropological principles of DA are thus adapted for the study of lead user identification with the aim of including unstructured big data for analysis. In this way, DA establishes the need of the study of the different sources of data, the need to understand the network of relationships among the data through triangulation and specific attention to each case study selected due to fact that culture is locally produced. In order to follow these principles, different social media are addressed (Facebook and YouTube), a triangulation of methods is followed to understand the relationships between the data, and a single case study is presented as the main focus of the current research.

4 | METHODOLOGY

This section presents the case study selected; the festival of Tomorrowland and the data collection. This research contains a triangulation of qualitative and quantitative methods. The sampling and data collection process is described for both types of data in the following sections.

4.1 | Case: Selection of the online community

Seeking to illustrate the conceptual argument that DA presents for lead users′ identification from open data, we focus on a case study (Siggelkow, 2007). The present case focuses on the music festival Tomorrowland (www.tomorrowland.com). The festival has been held in Belgium every year since 2005 and is considered as a part of the EDM (electronic dance music) movement, including styles such as trance, house, techno, breakbeat, gabber, and hardcore. The festival has gained success in attendance figures over the years, with tickets sold out in seconds upon release (Billboard, 2014). After each year′s edition, a promotional after-movie video lasting approximately 30 minutes is released on YouTube showing the best moments from the festival accompanied by the most relevant songs played by the DJs. The attractiveness of this specific case resides in several points. Firstly, the volume of data for analysis. Every year the after-movie on YouTube receives thousands of comments regarding the festival. These comments represent the interest of the community, being especially relevant for practitioners to understand how to use those insights and knowledge for service improvement. Secondly, the data is being produced continually; new comments are continuously being added even though the edition selected is the one from 2014 and other new videos from later editions are also available. Lastly, the data for this case comes from different sources with different potential for analysis (YouTube, Facebook, interviews, etc.). Hence, this case fulfills the conditions established as the "3Vs" (volume, velocity, variety) or the three characteristics of big data of McAfee and Brynjolfsson (2012).

4.2 | Tribal roles

Following the principles of DA, a closer study of the online community was performed in order to specify its characteristics. The

Tomorrowland community presents the structure of a "tribe". Tribes are relatively small groups of heterogeneous people bound by affection between their members (Cova & Cova, 2002). The special characteristics of this online community will serve to establish the different roles among its users that will serve to identify potential lead users in a rapid, cost-efficient, and unobtrusive manner. Cova and Cova (2002) establish that the members of these communities follow four types of roles: sympathizers, participants, practitioners, and devotees (see Figure 1).

In the case selected, the roles can be defined as follows: The group of **sympathizers** consists of members interested in what marketing shows about festivals, but their participation in those events may never happen, and their knowledge of things connected with the festival may be limited (Mitchell & Imrie, 2011). The group of **participants** is formed by those community members who are actively engaging in socialization with other members of the tribe; sometimes this interest is even stronger than their focus on the festival (Mitchell & Imrie, 2011). The group of **practitioners** is composed of those individuals who participate actively in the creation of the festival and who might have an economic interest in it. This group comprises DJs, managers, suppliers, web administrators, etc., who participate in the festival with an interest motivated by the meaning of the gathering (Mitchell & Imrie, 2011). The last group is the **devotees**, who are users with a high interest in this event, and who have a long-term engagement and knowledge about the tribe. This group considers the festival as a part of their identity, even when they do not assist physically.

4.3 | Digital anthropology as a method

Figure 2 explains the principles of DA as a flow chart and how these principles were applied to the case study of Tomorrowland. DA as a method follows five steps. The first step focuses on the selection of the case study that will be scrutinized, being in this case the EDM community of Tomorrowland. The community selected should contain different sources of data, as the second and third steps require the triangulation of the data gathered from those sources. In this particular case study, the community was situated around the social media of YouTube and Facebook. More specifically, the second step requires a closer look at these sources of data in order to establish whether there is some structure within the data that may contribute to the analysis of the big data sets. In our case, the community followed the structure of a tribe. This structure is found following the work of Cova and Cova (2002) in relation to the interviews and questionnaires carried out. The details of the interviews are disclosed in the next section on the qualitative data collection. The third step requires the selection of the methods necessary for the triangulation of the data collected. In our case study, this triangulation consisted of the mixture of qualitative and quantitative data as specified in the following sections. The fourth step seeks to find the relationships that exist in the data network. This step aims to understand the data holistically, one of the guiding principles of DA as discussed in the previous sections. In this case study, those relationships are gathered in Figure 3, where the relationships between the tribal roles of Cova and Cova (2002) and the lead user characteristics of Belz and Baumbach (2010) are triangulated and shown through spider web diagrams. Lastly, DA helps to identify the lead users in the selected community. In the current case study, this identification is done through the analysis of the big data set from YouTube over a content analysis of the data. This process is completed with the results of the research that stress how DA is more efficient than traditional lead user identification methods.

4.4 | Qualitative data collection

The qualitative data of this study comprises interviews, participant observation of the key respondents, and member checks. Additionally, a non-participant observation of two social media, YouTube and Facebook, was conducted during five months regarding the after-movie video of the festival. This observation regarded how relationships between members were performed as well as the topicality of the different themes presented in the threads of comments.

Two types of interviews were performed: three biographical unstructured interviews and 50 structured interviews through questionnaires sent via Facebook following a broadcasting process; the users that considered that their opinions could influence the future of the festival (self-assessment) filled in the online questionnaire. The three unstructured interviews (average length: 1.5 hours) were initially conducted to assess how members of the community would react to the study. This information was later used to prepare the

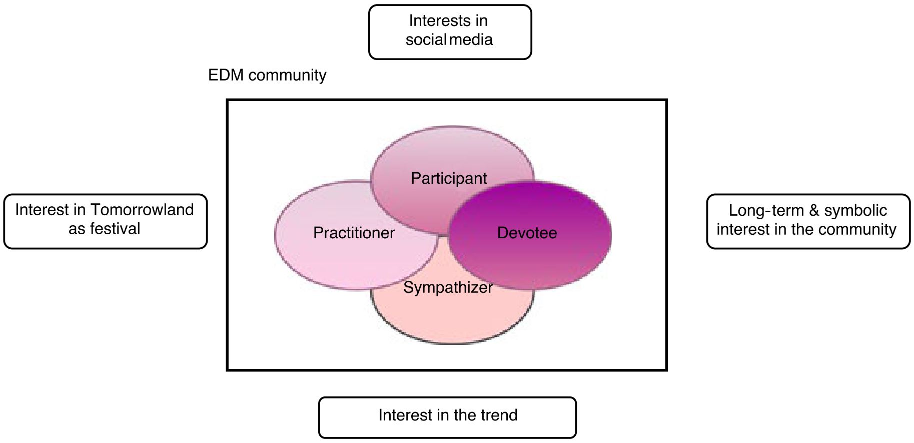
Interests in social media
EDM community
Interest in Tomorrowland as festival
Participant
Practitioner
Devotee
Sympathizer
Long-term & symbolic interest in the community
Interest in the trend

FIGURE 1 Character roles in the EDM community tribe (adapted from Cova & Cova, 2002) [Colour figure can be viewed at wileyonlinelibrary.com]

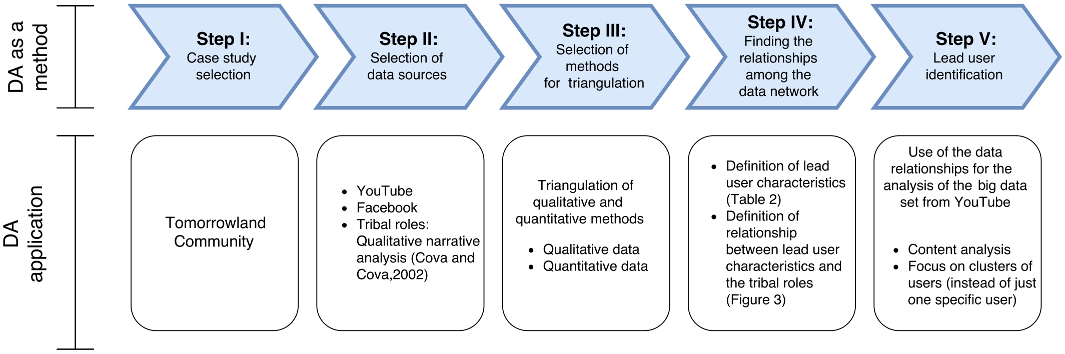
DA as a method
Step I: Case study selection
Step II: Selection of data sources
Step III: Selection of methods for triangulation
Step IV: Finding the relationships among the data network
Step V: Lead user identification
DA application
Tomorrowland Community
• YouTube
• Facebook
• Tribal roles: Qualitative narrative analysis (Cova and Cova,2002)
Triangulation of qualitative and quantitative methods
• Qualitative data
• Quantitative data
• Definition of lead user characteristics (Table 2)
• Definition of relationship between lead user characteristics and the tribal roles (Figure 3)
Use of the data relationships for the analysis of the big data set from YouTube
• Content analysis
• Focus on clusters of users (instead of just one specific user)

FIGURE 2 An overview of DA as a method and its application to the Tomorrowland case [Colour figure can be viewed at wileyonlinelibrary.com]

questionnaires looking for a suitable set of questions. The three individuals who engaged in an unstructured interview and shared their knowledge on EDM with the first author were considered key informants due to their inclination to be involved with the study (Bernard, 2011). Two of them were also considered to be specialists, due to their long-term relationship with EDM. All respondents were exposed to the same questions and the video. The set of questions was developed following the first unstructured interviews and can be found in Figure A1 in the Appendix.

Following the principles of DA, a closer study of the online community was performed in order to specify its characteristics. To this end, a qualitative study was performed using 10 respondents belonging to the community and one outsider, as a form of external validation of the sample. The outsider is an individual of the same age and characteristics as the rest of the group. This additional role is introduced in order to control whether the symbolic meanings accepted by the community members only belong to them, or whether they may be shared by others outside the community, thus being not specifically innovative and therefore lacking characteristics as lead users in this particular case. The qualitative data sample follows a nonprobability sample among users of the EDM festivals. The justification for this relies on Bernard's (2011) advice for studies based on one in-depth case study, informed by a long questionnaire seeking deep understanding of the respondents. Bernard's (2011) work regarding nonprobability sampling establishes that "there is growing evidence that 10–20 knowledgeable people are enough to uncover and understand the core categories in any well-defined cultural domain or study of lived experience" (Bernard, 2011, p. 154).

These respondents were found offline or in social media other than YouTube. This was done in observance of the principles of DA, which prescribes the use of a different source of data for triangulation. The respondents willing to participate with member checks of the study were two males (from Spain and Denmark) and two females (from Spain and Hungary), aged between 25 and 31 years old. The rest of the respondents were in a range between 16 and 35 years old. The rest of the respondents were not especially interested in performing member checks or in the outcome of the study. The member checks consisted in a cyclical process in which the author went back and forth to the respondents who had previously been presented with her assertions for analysis by the community.

4.5 | Quantitative data collection

The quantitative data is constituted by all retrievable YouTube users' comments in response to the after-movie during the period September 16, 2014, to January 16, 2015. During this period the number of comments on the video on YouTube reached a total of 9,149, of which 2,328 where conversations among users (i.e. comments with at least one reply). The process to create the "big data" set consisted of extracting all the possible comments that the html allowed on the YouTube webpage. Once the maximum number of comments was gathered in a PDF document, MATLAB Simulink was used to extract the information available from the comments and users on YouTube. This information was coded as shown in Table 1.

5 | ANALYSIS

We followed the principles of DA to find the relationships among the whole network of data (Step IV in Figure 2). A stepwise procedure was followed in order to produce a holistic analysis from the YouTube thread of comments, as explained in the following.

5.1 | Lead users' characteristics and tribal roles

In this section we compare the characteristics proposed by Belz and Baumbach (2010) to identify lead users ("ahead of trend", "dissatisfaction", "product-related knowledge", "use experience", "involvement", and "opinion leadership") with the tribal roles of Cova and Cova (2002) (Devotee, Practitioner, Participant, and Sympathizer). In the first place, an example of the characteristics of Belz and Baumbach (2010) is presented in Table 2 to show how the different characteristics were defined in the study.

The different types of interviews and questionnaires performed helped to define the different roles of Cova and Cova (2002) in our case study. Table 3 presents an example of how the different roles were defined following the answers of the respondents. Additionally, a core interests for each role was defined.

The relationship between the tribal roles and the lead user attributes defined by Belz and Baumbach (2010) and Lüthje (2004) can be observed in Figure 3. The intensity of each characteristic was defined for each role regarding its intensity in this order: from the least

TABLE 1 Coding system for the classification of the users′ comments

Code	Content
ID	Order of the comment in the thread of comments
User	Name of the user posting a comment
Conversation	Situation of the post in the thread of comments
Number of replies	Number of replies to a comment
Replies?	This tag stated whether the comment was an original comment or a reply
Text	Textual content of the post
Likes	Number of likes that the post achieved on the thread

TABLE 2 Characteristics of lead users and exemplary statements of the EDM community members indicating lead user attributes

Lead user characteristics	Typical examples
Ahead of trend	"The workers of Tomorrowland are like members of the cast, entertaining you all the time, they wear costumes with flowers, bubbles, etc. You would not see that in Denmark, the effort to make it like a fairy tale. They should do it that way […] You are entertained from the beginning to when you leave and you can keep the fairy tale spirit in people."
Dissatisfaction	"I hate the term 'EDM'. Its kinda the new word for electronic house, but it′s like a lot genres have been generalized to just this term 'EDM'. Big room, progressive house, electro, bounce and some other genres got pushed in there, and it′s such a mainstream word for those genres."
	"I can′t wait to go until I get old enough", the 18 year restriction is something wrong."
Use experience	"I went to Tomorrowland in 2012 and 2014, there are some local festivals but nothing compared to it. Tomorrowland is the biggest there is, has all my favorite music and DJs, people go there and have fun together from all around the world, the organizers make a great job in setting up the place with lots of imagination and cool stuff."
Product knowledge	"Tomorrowland started in Belgium."
	"Much of the lyrics are meaningless on the pop music […] But EDM express some kind of philosophy in each of the songs […] Armin van Buuren dedicated one of his tracks to his newborn son and there is a description of a lot of his feelings at that time, something that matters. Beyoncé says: 'who runs the world', ok, 'girls', and then she repeats the same 14 times and then the song is over."
	"Tomorrowland is also the most expensive festival I ever heard of. They usually last 3 days and they are smaller and have less known DJs, so they are cheaper … Everything goes between 600kr to 1500kr and that is a bit more affordable."
Involvement	"I went by myself from India […] The festival in itself was amazing. I never felt that I was alone, and even when I was walking around looking a bit confused trying to find different stages, I had loads of people ask me if I was ok and if I needed help."
Opinion leadership	"Yeah of course I talk about [Tomorrowland] with my friends who also have been there and we compare it with the after-movies from other years. And when someone asks me about Tomorrowland I suggest they watch the after-movie."

TABLE 3 Tribal roles in the EDM community

Role	Example	Role interests
Sympathizer	What is your favorite sentence from Tomorrowland and why?	Marketing image
	"The ones that show nice summer images and beautiful girls while listening to good songs..."	
Participant	Have you ever discuss about something related to the festival with other people in any social media?	Socialization
	"Yeah often on Facebook"	
	Have you ever talked about this video with someone?	
	"Yes, of course, I talk about it with my friends […] and we compare the after-movies from other years. And when someone asks me about Tomorrowland I suggest they watch the after-movie."	
Practitioner	"I went by myself from India […] The festival in itself was amazing. I never felt that I was alone, and even when I was walking around looking a bit confused trying to find different stages, I had loads of people ask me if I was ok and if I needed help."	The festival
Devotee	How did Tomorrowland start? "It started like an ordinary festival, the first years it was David Guetta and a few less known DJs, and then there were 5,000 people and it grew bigger and bigger and bigger. If you watch the after-movie of 2012, David Guetta is saying that he′s been there since the first time and that they are getting bigger each year and he is amazed how many people music can bring together despite religion and flags ... it is just music."	Long-term relationship with the community and the festival

relevant within the roles (4) to the most relevant attribute (1). Table A1 in the Appendix shows the justification of this coding on each characteristic for the different roles.

Figure 3 shows how the different roles are characterized by the different attributes of Belz and Baumbach (2010). The **Devotees** are characterized by being the leaders regarding the prospect trends

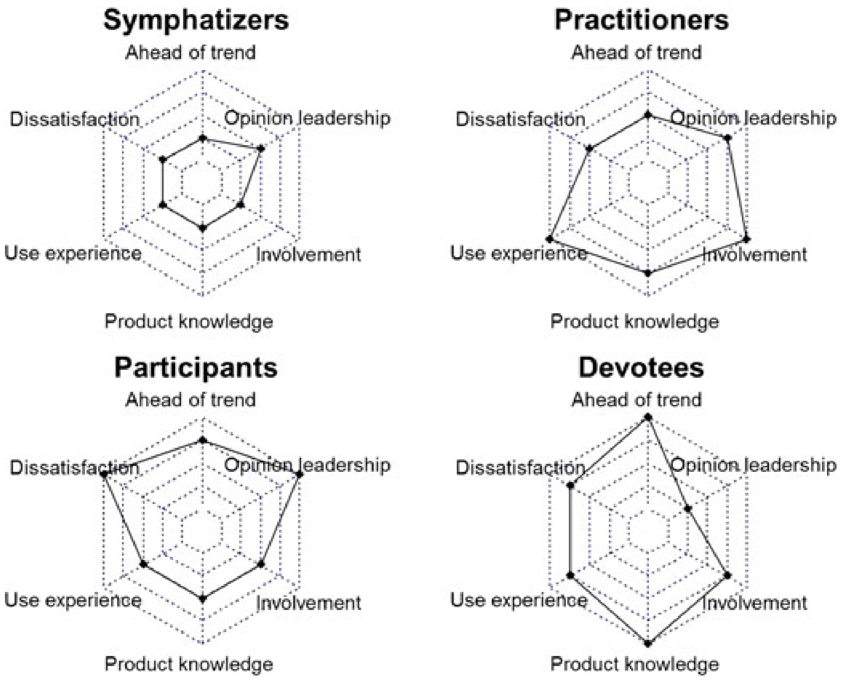

FIGURE 3 Relationship between the tribal roles and the lead users′ characteristics on the data [Colour figure can be viewed at wileyonlinelibrary.com]

(ahead of trend) and the product knowledge derived from years of experience within the festival sector. The **practitioners** are characterized by having an extended use experience of the Tomorrowland festival; their experience with the industry might be shorter than the experience of the devotees; nevertheless, they have a deeper engagement with this festival specifically and thus they have a bigger involvement. Due to their special interest in the social aspect of the festival, and more specifically, on the social media channel, the **Participants** are the opinion leaders. They often share meaning and opinions of the festival online. Those opinions are often to voice their dissatisfaction with the music and to debate other issues such as the age entry requirements for the festival. Due to their relatively new and superficial interest in the festival and the industry, the **Sympathizers** are situated as the least interesting role regarding the characteristics studied, nevertheless sharing a general understanding of the festival and an interest in the socialization aspect that is portrayed by the marketing image of the festival.

5.2 | Roles in the big data set

In this section, the different characteristics and roles will be used to study the user types that may be found in the YouTube big data set following the principles of DA, thus addressing the structure of the data source of our case study. In order to do this, first, an analysis of the thread of comments will be produced attending to the coding presented in the analysis section (Step V in Figure 2). In this regard, seven variables were constructed from the data structure of the comments. The conversations were used as a sign of interest in a comment, triangulating that interest with the number of likes attributed to the conversation. Additionally, a temporal factor was taken into account, attending to the moment in which the comment appeared in the thread. This is done for the seven variables established in Table 1. YouTube orders the comments according to several criteria, such as the number of likes that a comment has, for example. We use this variable to order the comments regarding the time in which they were

TABLE 4 Big data set themes and examples

Theme	Sample comment	Statistical weight
Gaining popularity	*“OK who wants to go next year???”*	505 replies 21.7% of the conversations 1250 likes
Information about the video/songs	*“00:50 ID - ID vs Sander Van Doorn, Martin Garrix, DVBBS – Gold Skies (feat. Aleesia)* *02:50 Dimitri Vangelis & Wyman X Steve Angello – Payback* *04:04 Yves V & Don Diablo – King Cobra* *… … …”*	175 replies 7.51% of the conversations 1093 likes
Dissatisfaction	*“This world is so messed up. Humans are dying every day from diseases and hungry and others are partying like it′s what this life is about. But when you think deeply, you find that those who are suffering understood the meaning of life, and didn′t accept to be treated badly from a small tiny minority of people and others are the victims.”*	110 replies 4.72% of the conversations 25 likes
Humorous content	*“Hello, I am currently 17 years old and I want to become a walrus. I know there′s a million people out there just like me, but I promise you I′m different. On January 14th, I′m moving to the Arctic. I′ve already cut off my arms, and now slide on my stomach everywhere. I write with my teeth. I may not be a walrus yet, but I promise you if you give me the support I need by subscribing to my channel, I will become the greatest walrus ever.”*	62 replies 2.66% of the conversations 359 likes

posted. Table 4 explains the themes found in the big data set and reports the statistical weight of each theme.

The first theme represents an interest in socialization from the users. This theme is the most relevant appearing in 21.7% of the conversations. The second theme corresponds to product knowledge, as the interviews with the devotee showed; this group uses the YouTube videos as a playlist and the comments as a source to find information regarding the video, such as a list of the songs played. The third theme, dissatisfaction, is closely related to the socialization interest due to the fact that it is used to share users′ opinions regarding what does not work in the festival. The last theme can also be considered as a socialization attribute due to its humorous content. Taking into account these results, and the fact that the theme with the biggest weight in the data set is socialization, we determine that the social media channel is marked primarily by the participant role. In the Facebook channel, a broadcasting process was followed; the problem seek was presented to different groups and 50 questionnaires were filled by those users that deemed themselves as having the right answers to help develop the future of the festival. This group was characterized by users sharing their use experience of Tomorrowland and how they were involved with it to the point of going along to the festival or working just for the sake of being able to afford to go to the festival. Following the same reasoning as before, they were classified as practitioners due to the attributes presented by this group of users (use experience and involvement).

Following the principles of DA, the different social media channels were classified according to the different aspects that the users presented in each of them (the tribal roles of Cova & Cova, 2002). The results obtained demonstrate that the different architecture of the channels impacts the content that the users may produce and thus attracts different user types. YouTube has as its main focus the video, with portraits appearing on it as the final socializing end; in this sense, it is reasonable that users with a high focus on socialization (the participants) are attracted to this channel. The Facebook groups did not contain as many comments as the YouTube channel; however, their content was more extensive, thus being preferred by those members that wished to explain deeper insights in their comments. In this case, these users aimed to explain their involvement and experience in the anchoring place of the festival, being thus classified as practitioners. This terminology refers to the actual place in which the festival is celebrated; that Cova and Cova (2002) establish as the anchor that holds the tribal community together. The last two roles, sympathizers and devotees, were not as present in the data. Due to their shallow knowledge of the festival, the sympathizers were not inclined to comment in any of the channels studied as the member checks revealed. Due to their lack of interest in the online socialization process, the devotees were not so easily found in the data. The member checks demonstrated in this case how this role preferred face-to-face meetings with other devotees with whom they had a friendship based on their assistance at different festivals throughout Europe, not just specifically Tomorrowland. This role was found following a snowballing technique with users that had a face-to-face interview, asking them to propose other respondents that they assessed as fans of the EDM music or the Tomorrowland festival. It is worth mentioning how the devotees in this study did not classify themselves as members of this group, being categorized as such by the researchers based on the interviews. These results highlighted the problems with self-assessment for lead user identification.

6 | DISCUSSION

The application of DA to the network of relationships among the data in this case study revealed the presence of the four roles of the tribal community from the festival and helped to establish which characteristics were the main attributes of each different role. Four themes of importance to the online community were identified. Using the roles stated by Cova and Cova (2002) and the lead user indicators of Belz and Baumbach (2010) (ahead of trend, dissatisfaction, product knowledge, use experience, involvement), our analysis in the context of open data showed that different social media channels are composed by users with different lead characteristics. As mentioned at the beginning of the paper, the lead user identification methods presented in the literature have as principal focus the individual and its capacity as lead user and therefore as producer of knowledge (Stockstrom et al., 2016).

6.1 | Theoretical contributions

The results of this study require a shift from the usual concepts in the literature in order to adapt and make possible the analysis of data in a data-rich environment (Bosch-Sijtsema & Bosch, 2015). The study of big data requires a change of focus from specific individuals to a group of individuals with the capacity to produce the insights of a lead user. Big data implies the extension of analysis to whole populations instead of specific samples, bringing about new possibilities to model and extract the insights of consumers publicly exposed via digital media. In this sense, DA focuses on the different possibilities for innovation that each of the roles may produce, revealing the importance all the roles present in the case study. This result contrasts with netnography, which regards the devotees as the main source of insights, being thus the main focus of that method in the search for this specific type of user.

The main attractiveness of DA is the possibility of studying a case study in a detailed manner, taking into account the different specificities of the data found in the different sources. Furthermore, through triangulation, DA avoids over-reliance on user self-assessments, and is thus advantageous compared to screening and pyramiding.

Nevertheless, using a new method like DA also presents some challenges. A limitation of this study may come from the coding of the big data set. Due to its unstructured nature (text), we pursued a content analysis of the data to establish the different themes that are important for the community. In this case, we considered the conversations as a signal of interest for the members of the community. In future applications, we suggest following different coding possibilities attending to the structure of the data in order to attain higher validity. There is a degree of subjectivity involved in the processing of coding the data, and thus errors regarding the interpretation of the users′ comments may occur and should be reduced, for example by employing independent coders.

Other key limitations are the difficulty of generalizing this approach from case to case without an exhaustive study of the specific data structure, and the fact that the analysis of the data is time consuming.

This paper has focused on users from the entertainment industry, but more case studies focusing on different products and services are advised to further verify the suitability of the approach presented. Furthermore, the combination of methods presented in this document is flexible with regard to which methods can be used for the triangulation of the results. A selection of some key devotees can be the base of a netnographic study that can be set for comparison of the results that the whole data set produces in combination as presented in this paper. Additionally, a comparison of lead user identification from DA with those identified through traditional off-line techniques should be considered in future work.

6.2 | Managerial implications

Mass screening is currently the most common approach for lead identification (Belz & Baumbach, 2010). As mentioned earlier, this method presents issues such as reliance on self-assessment, high costs, and correct sampling. The new possibilities offered by open data can potentially be used to solve these problems: self-assessment is substituted by triangulation as shown in the empirical case study, the cost of gathering the data is limited due to its public availability on the internet, and big data tools allow the whole population subject of study to be easily mapped. In this sense, DA combined with big data analysis is a good and efficient option for practitioners to find lead users in an inexpensive manner.

Methods like pyramiding, screening, and broadcasting are simply means for identifying lead users: All three methods aim to identify but not to capitalize on the resources that reside with the lead users. A substantial managerial advantage of DA in comparison to these methods is that DA is a method for *both* identifying *and* capitalizing from the resources of the lead users. Hence, as part of the DA identification, the searcher (e.g., a company) can readily tap into the reflections and suggested solutions presented by the identified lead users.

DA combines some of the advantages originally addressed by netnography (Belz & Baumbach, 2010) and at the same time solves some of its issues, thus posing a natural next step forward for researchers and companies. The main drawback of DA is constituted by the fact that studying the structure in the thread of data is time consuming, especially the first time it is conducted. Nevertheless, knowing this structure may bring many other benefits to companies, e.g. knowledge of the most relevant themes in which users are interested. This can be used for keyword research and search engine marketing (SEM) in order to improve visibility in the search engines, thereby reducing cost in the marketing department.

Another advantage of the perspective of DA is the fact that there is no need to find a unique user that has all lead user characteristics as proposed by Belz and Baumbach (2010). This represents an improvement in terms of efficiency compared to netnography. Finally, DA, as a natural step after netnography, brings some of its benefits, such as the dissatisfaction of the users regarding current product on the market as well as the explanation of the consumer's pain, that can be used for service innovation and improvement (Belz & Baumbach, 2010; Kozinets, 2002, 2010). Relationships with the users are not a specific need outside of the control group, thus a bigger quantity of ideas are collected, with also the possibility of a thick content for those ideas triangulating the data collection from different sources. For all these reasons, the application of DA to unstructured big data may represent a good opportunity for companies to be ahead of competitors and in tune with their customers.

ORCID

Vella Verónica Somoza Sánchez http://orcid.org/0000-0003-0928-2377
Davide Giacalone http://orcid.org/0000-0003-2498-0632
René Chester Goduscheit http://orcid.org/0000-0001-8639-2014

REFERENCES

Belz, F. M., & Baumbach, W. (2010). Netnography as a method of lead user identification. *Creativity and Innovation Management*, *19*, 304–313.

Bernard, H. R. (2011). *Research methods in anthropology: Qualitative and quantitative approaches* (5th ed.). Lanham, MD: AltaMira Press.

Bharadwaj, N., & Noble, C. H. (2015). Innovation in data-rich environments. *Journal of Product Innovation Management*, *32*, 476–478.

Billboard. (2014). Tomorrowland sells 360,000 tickets in under an hour. Retrieved on 23 March 2017 from http://www.billboard.com/articles/columns/code/5908407/tomorrowland-sells-360000-tickets-in-under-an-hour.

Boellstorff, T. (2012). Rethinking digital anthropology. In H. A. Horst, & D. Miller (Eds.), *Digital anthropology* (pp. 39–60). Oxford: Berg Publishers.

Bosch-Sijtsema, P., & Bosch, J. (2015). User involvement throughout the innovation process in high-tech industries. *Journal of Product Innovation Management*, *32*, 793–807.

Brem, A., & Bilgram, V. (2015). The search for innovative partners in co-creation: Identifying lead users in social media through netnography and crowdsourcing. *Journal of Engineering and Technology Management*, *37*, 40–51.

Chatterji, A. K., & Fabrizio, K. R. (2014). Using users: When does external knowledge enhance corporate product innovation? *Strategic Management Journal*, *35*, 1427–1445.

Cohen, W. M., Nelson, R. R., & Walsh, J. P. (2002). Links and impacts: The influence of public research on industrial R&D. *Management Science*, *48*, 1–23.

Cova, B., & Cova, V. (2002). Tribal marketing: The tribalisation of society and its impact on the conduct of marketing. *European Journal of Marketing*, *36*, 595–620.

Feldman, M., Kenney, M., & Lissoni, F. (2015). The new data frontier: Special issue of *Research Policy*. *Research Policy*, *44*, 1629–1632.

Jeppesen, L. B., & Lakhani, K. R. (2010). Marginality and problem-solving effectiveness in broadcast search. *Organization Science*, *21*, 1016–1033.

Kozinets, R. V. (2002). The field behind the screen: Using netnography for marketing research in online communities. *Journal of Marketing Research*, *39*, 61–72.

Kozinets, R. V. (2010). *Netnography: Doing ethnographic research online*. London: Sage.

Kratzer, J., & Lettl, C. (2009). Distinctive roles of lead users and opinion leaders in the social networks of schoolchildren. *Journal of Consumer Research*, *36*, 646–659.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*, 1121–1134.

Laursen, K., & Salter, A. (2006). Open for innovation: The role of openness in explaining innovation performance among UK manufacturing firms. *Strategic Management Journal*, *27*, 131–150.

Lüthje, C. (2004). Characteristics of innovating users in a consumer goods field: An empirical study of sport-related product consumers. *Technovation*, *24*, 683–695.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Eamon Dolan/Mariner Books.

McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*, *90*, 60–68.

Miller, D., & Horst, H. A. (2012). The digital and the human: A prospectus for digital anthropology. In H. A. Horst, & D. Miller (Eds.), *Digital anthropology* (pp. 3–35). Oxford: Berg Publishers.

Mitchell, C., & Imrie, B. C. (2011). Consumer tribes: Membership, consumption and building loyalty. *Asia Pacific Journal of Marketing and Logistics*, *23*, 39–56.

Murugesan, S. (2009). *Handbook of research on Web 2.0, 3.0, and X.0: Technologies, business, and social applications*. Hershey, PA: IGI Global.

Poetz, M. K., & Prügl, R. (2010). Crossing domain-specific boundaries in search of innovation: Exploring the potential of pyramiding. *Journal of Product Innovation Management*, *27*, 897–914.

Prügl, R., & Schreier, M. (2006). Learning from leading-edge customers at *The Sims*: Opening up the innovation process using toolkits. *R&D Management*, *36*, 237–250.

Sawhney, M., Verona, G., & Prandelli, E. (2005). Collaborating to create: The internet as a platform for customer engagement in product innovation. *Journal of Interactive Marketing*, *19*, 4–17.

Siggelkow, N. (2007). Persuasion with case studies. *Academy of Management Journal*, *50*, 20–24.

Stockstrom, C. S., Goduscheit, R. C., Lüthje, C., & Jørgensen, J. H. (2016). Identifying valuable users as informants for innovation processes: Comparing the search efficiency of pyramiding and screening. *Research Policy*, *45*, 507–516.

Sudman, S. (1985). Experiments in the measurement of the size of social networks. *Social Networks*, *7*, 127–151.

von Hippel, E. (1986). Lead users: A source of novel product concepts. *Management Science*, *32*, 791–805.

von Hippel, E. (1988). *The source of innovation*. New York: Oxford University Press.

von Hippel, E., Franke, N., & Prügl, R. (2009). Pyramiding: Efficient search for rare subjects. *Research Policy*, *38*, 1397–1406.

von Hippel, E., Thomke, S., & Sonnack, M. (1999). Creating breakthroughs at 3M. *Harvard Business Review*, *77*, 47–57.

Vella V. Somoza Sánchez is a PhD student at the Center of Integrative Innovation Management at the University of Southern Denmark (SDU). Her primary research interests focus on two major areas. First, she investigates and develops new methodologies for Big Data analysis using combinations of structured and unstructured data. Second, she focuses on service innovation and organizational communication strategies through leadership.

Davide Giacalone is an associate professor in the Department of Technology and Innovation, University of Southern Denmark. His research centers on consumers′ sensory perceptions and acceptance of product innovations. His areas of ongoing research include consumer acceptance of novel products and technologies, sensometrics, in particular multivariate modeling of sensory and consumer data, Big Data mining and its applications in product development and innovation.

René Chester Goduscheit is an associate professor in Innovation Management at the University of Southern Denmark. His primary research interest focuses on open innovation and external sources of innovation. His work covers research, action research, consultancy, and evaluation of projects/programs aimed at innovation. He currently works with a long list of larger corporations, small and medium-sized and public organizations in a common effort to get insights into (open) innovation management.

How to cite this article: Somoza Sánchez VV, Giacalone D, Goduscheit RC. Digital anthropology as method for lead user identification from unstructured big data. *Creat Innov Manag*. 2017;1–10. https://doi.org/10.1111/caim.12228