

Social Media-Based Collaborative Information Access: Analysis of Online Crisis-Related Twitter Conversations

Lynda Tamine, Laure Soulier, Lamjed Ben Jabeur, Frédéric Amblard, Chihab Hanachi, Gilles Hubert, Camille Roth

▶ **To cite this version:**

Lynda Tamine, Laure Soulier, Lamjed Ben Jabeur, Frédéric Amblard, Chihab Hanachi, et al.. Social Media-Based Collaborative Information Access: Analysis of Online Crisis-Related Twitter Conversations. 27th ACM Conference on Hypertext and Social Media (HT 2016), ACM: Association for Computing Machinery, Jul 2016, Halifax, Nova Scotia, Canada. pp.159 - 168, 10.1145/2914586.2914589 . hal-03597237

HAL Id: hal-03597237

https://sciencespo.hal.science/hal-03597237

Submitted on 4 Mar 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

Social Media-Based Collaborative Information Access: Analysis of Online Crisis-Related Twitter Conversations

Lynda Tamine
IRIT, University of Toulouse, UPS
118 Route de Narbonne
Toulouse, France
tamine@irit.fr

Laure Soulier
Sorbonne Universités, UPMC Univ Paris 06, CNRS UMR 7606, LIP6
F-75005 Paris, France
laure.soulier@lip6.fr

Lamjed Ben Jabeur
IRIT, University of Toulouse 3, UPS
118 Route de Narbonne
31062 Toulouse, France
jabeur@irit.fr

Frederic Amblard
IRIT, University of Toulouse 1 Capitole
2 rue du Doyen Gabriel Marty
31042 Toulouse Cedex, France
amblard@irit.fr

Chihab Hanachi
IRIT, University of Toulouse 1 Capitole
2 rue du Doyen Gabriel Marty
31042 Toulouse Cedex, France
hanachi@irit.fr

Gilles Hubert
IRIT, University of Toulouse, UPS
118 Route de Narbonne
Toulouse, France
hubert@irit.fr

Camille Roth
Centre Marc Bloch Berlin, UMIFRE CNRS-MAE
Berlin, Germany
roth@cmb.hu-berlin.de

ABSTRACT

The notion of implicit (or explicit) collaborative information access refers to systems and practices allowing a group of users to unintentionally (respectively intentionally) seek, share and retrieve information to achieve similar (respectively shared) information-related goals. Despite an increasing adoption in social environments, collaboration behavior in information seeking and retrieval is mainly limited to small-sized groups, generally restricted to working spaces. Much remains to be learned about collaborative information seeking within open web social spaces. This paper is an attempt to better understand either implicit or explicit collaboration by studying Twitter, one of the most popular and widely used social networks. We study in particular the complex intertwinement of human interactions induced by both collaboration and social networking. We empirically explore explicit collaborative interactions based on focused conversation streams during two crisis. We identify structural patterns of temporally representative conversation subgraphs and represent their topics using Latent Dirichlet Allocation (LDA) modeling. Our main findings suggest that: *1)* the *critical mass* of collaboration is generally limited to small-sized flat networks, with or without an influential user, *2)* users are active as members of weakly overlapping groups and engage in numerous collaborative search and sharing tasks dealing with different topics, and *3)* collaborative group ties evolve within the time-span of conversations.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

HT '16, July 10-13, 2016, Halifax, NS, Canada

© 2016 ACM. ISBN 978-1-4503-4247-6/16/07…$15.00

DOI: http://dx.doi.org/10.1145/2914586.2914589

CCS Concepts

•**Information systems** → **Collaborative and social computing systems and tools;** *Information systems applications;* Social networking sites;

Keywords

Information Access; Social Networks; Twitter; Topic Models; Collaboration

1. INTRODUCTION

Using social networking platforms for information seeking and sharing is an increasingly common practice [28]. Although previous research [7] and various services, such as Aardvark [20], have investigated the use of social media for information access, the underlying paradigm still relies on individual search. In this setting, the information access is generally enriched by cues stemming from a seeker's social relationships (e.g., so-called "friends" or "followers"). However, recent studies highlighted the fact that a significant portion of information searches remains unsolved within the user's social neighborhood [23]. To address this issue, we believe that search engines could support the creation of social ties between users or groups of users aiming at carrying out similar search tasks. The long-term goal is to favor both explicit and implicit collaboration. In explicit collaborative search scenarios, two or more individuals (say, a work team) are engaged in the search process and intentionally combine their knowledge and skills to solve a shared information need. In contrast, implicit collaboration refers to scenarios where users might unintentionally share their experience with other users to satisfy their own information needs [16].

A sizable literature has shown that explicit collaboration within small-sized groups is beneficial to information search [36]. Yet, a research topic which is still under exploration

deals with the opportunities of large-scale explicit collaboration supported by social networking platforms [30, 32]. We aim to contribute to this emerging topic by studying the properties of groups of users with shared interests emerging from online social conversations, viewed here as collaboration signals. We empirically explore explicit collaborative interactions based on focused conversation streams during two crisis. We identify structural patterns of temporally representative conversation subgraphs and represent their topics using Latent Dirichlet Allocation (LDA) modeling. We focus on a crisis management scenario based on two Twitter-based datasets collected during critical circumstances (Hurricane Sandy[1] and Ebola[2]). The reasons for choosing specifically this scenario are twofold: *1)* social media platforms are increasingly used by citizens during crisis situations [21] to make helpful information and knowledge available (events, video, expertise, etc.), to ease crises awareness, to accept or distribute tasks to volunteers, and to share their opinions on the way crises are managed by official responders, *2)* crisis-related situations lead to the emergence of spontaneous groups of users willing to collaborate (through online communications) in order to provide resources and help victims [18].

Our main contributions include:

- Characterizing both the structural and semantic patterns of explicit collaboration, based on information seeking and sharing traces as well as signals left by groups of users who are jointly engaged in temporally tight conversations.
- Exploring whether and how much groups of users with similar interests may be more likely to explicitly collaborate with each other.

Our findings provide a number of significant opportunities for future research on collaboration in the social Web. More precisely, they may guide designers of social media-based information search tools to connect users to a wide and relevant audience in accordance with their search goals and interests

The remainder of this paper is organized as follows. In section 2, we give an overview of the relevant literature on social and collaborative information search, and then focus on the main challenges and research advances in social-media based crisis management. Section 3 details the data acquisition and processing methods. In section 4, we present and discuss the results. Section 5 highlights the broader implications of our study in terms of social-media collaboration resarch, while section 6 outlines its limitations and future research directions. Some concluding remarks follow in section 7.

2. BACKGROUND AND RELATED WORK

This paper is related to two lines of previous work that we overview. We first focus on the social and collaborative information access using social media platforms and then, investigate the use of social media services during crisis situations.

[1]https://en.wikipedia.org/wiki/Hurricane_Sandy
[2]https://en.wikipedia.org/wiki/West_African_Ebola_virus_epidemic

2.1 Collaborative and social media information seeking: two sides of the same coin?

Although generally perceived as a solitary process, information seeking and retrieval increasingly imply collaboration with others either within small-sized work teams [36] or open social web spaces [30]. The first research initiative dealing with the use of social media and favoring large-scale collaboration has been raised by the DARPA challenge aiming at identifying ten red balloons across the USA [40]. Collaboration could be defined according to various dimensions: namely, intent, depth, concurrency, and location, leading to fundamentally distinct processes (e.g., recommendantion, task-based search) and research challenges [16, 33]. With respect to the above-mentioned objectives, we focus in this paper on the intent dimension which can be either explicit or implicit.

Explicit collaboration has commonly been addressed in the area of collaborative information seeking and retrieval [14, 36]. In this context, an important paradigm for the optimization of the collaborators' search actions is the *division of labor*. This could be traced to three types of mediation which, in turn, correspond to specific user behaviors: *1)* user-based mediation through explicit discussions or exchanges between the collaborators using interfaces [36], *2)* system-based mediation supported by search algorithms which transfer the results to the right collaborators, generally according to their predefined roles [38], *3)* hybrid mediation that learns and assigns evolving roles to users and adapts the search accordingly [37]. Implicit collaboration has traditionally been addressed in algorithms and applications such as collaborative filtering [6].

Recently, social media platforms have given rise to both explicit and implicit collaborative search under the umbrella of "social search" [13, 29, 30]. Authors in [29] broadly define social search as *"the process of finding information online with the assistance of other social resources as well as search over collections of socially-generated content"*. While some works exploit social signals (*like/dislike*, comments) [2] or social features (engagement, trust) [24] to enhance implicit collaborative search models, other research strands closer to ours focusing on how users appropriate social media platforms to explicitly collaborate. The main findings may be subsumed as follows: *1)* seeking and sharing information are the two basic forms of online explicit collaboration using social networks [10, 29], *2)* the main motivations of users to explicitly ask their social network are trust, awareness (social support), searching for opinionated information and reaching specific audience [29, 23] and that *3)* a significant part of the questions asked to social networks did not receive answers mostly because of the low social activity of askers or the limited size of their social neighborhood [23].

In this paper, we reveal some characteristics about the behavioral facets of users engaged in social-media based explicit collaboration. Unlike previous work [9, 10, 23, 29], our study specifically focuses on: *1)* the characterization of the group patterns of users engaged in explicit collaboration supported by topically focused online conversations that express shared information goals and interests; *2)* the examination of the social connectivity between these groups in order to have a picture of their interplay regarding both the generation of content and the social interactions. In addition, different from the study presented in [9], our study focuses on the group level rather than the user level. More

specifically, rather than aiming to discover the user-to-user interaction graph structure regardless of the users' intent, our goal here is to give an abstracted picture of the collaborative group patterns that emerge from the user-to-user social interactions regarding shared information needs and interests.

2.2 Social-media based collaboration in emergency situations

Besides conventional social media (e.g., Facebook, Twitter, etc.) that are commonly used in crisis management, new dedicated platforms have recently emerged (Sahana, Ushahidi, OneResponse, Tweet4Act, Google Crisis Response, etc.). Some tools (i.e., NYPA) offer several functions while interfacing with conventional tools and Geographic Information Systems. In terms of use, these tools allow *1)* citizens to geo-locate elements (events, victims, demands, resources, etc.) in order to be informed and active in resolving the crisis as well as *2)* stakeholder organizations to collaborate and be more efficient, which results in accelerating decision making and therefore action. In this context, Twitter, Facebook, and Ushahidi are the most used social platforms in crisis situations notably during the earthquake in Haiti in 2010 and the Fukushima nuclear accident in 2011 [42]. A detailed review of social-media processing in crisis-management can be found in [21].

Based first and foremost on Twitter streams, a large amount of research studies proposed approaches to predict crises [19, 26] and model information spread [35]. This enables to improve communication channels or understand users' individual and collective behaviors that highlight situational awareness [1, 17, 41]. Heverin and Zach [17] highlighted the collective effort made by users to produce information that could be identified within "the chaos" through the hashtag stream, aiming at forging a global picture of the overall crisis-related information. Vieweg et al. [41] distinguished among various content-based tweet features those which could be used to identify ad hoc audiences viewed as potential collaborators. The topical analysis of tweets showed that different types of information may be broadcasted depending on the role of those seeking information. For instance, *Preparation* and *Response to warning* concern both individual and organizational audiences. However confidence-related challenges remain, especially with respect to the reliability of both partners' commitment and shared information. On the whole, they limit the use of social-media based technologies to enhance citizen-to-organization and organization-to-organization collaboration [25].

Our work extends previous work [17, 41] about users' involvement in crisis-related social-media streams by giving a picture of the structure of the *intra* and *inter* collaboration between user groups with the aim of highlighting collaboration opportunities.

3. STUDY DESIGN

3.1 Definitions and assumptions

We introduce here the basic definitions and assumptions:

- *Collaboration.* Collaboration allows people to create and share collective knowledge within a work team to identify and solve a shared complex problem [34]. From the social web point of view, the collaboration concept is closely related to the notion of wisdom of crowds, i.e., how large groups of people are better at solving problems and fostering innovation [39].

- *Collaboration signals in the social web.* To detect collaboration intent of active users in the social web, we use the following assumption:

  Assumption 1. *Providing assistance to others by means of social signals like answering, sharing, and propagating information through the social network is a form of online collaboration [13].*

- *Collaborative information search and sharing task.* We consider here that a collaborative search and sharing task is implicitly performed through a conversation started by a seed tweet, whereby other participants are trying to address the same issue. All of them strive to achieve a common task that consists, for instance, in the retrieval of a specific information or the synchronization around a particular action. Accordingly, we will later consider in our study the following assumption:

  Assumption 2. *Online conversations are timely bounded and could convey different subtopics and subtasks alongside their lifetime [9].*

- *Collaborative group.* We consider the active users involved in a conversation as the members of the collaborative group.

3.2 Twitter Datasets

We selected two datasets obtained by constantly monitoring Twitter's stream via Twitter Streaming API using appropriate tracking keywords during a critical period of two different crises. In particular, we used the *filter* method of streaming API that delivers tweets which imperatively contain tracked keywords. The *filter* method enables to gather a higher number of tweets about the monitored crises in comparison to the *sample* method that randomly pushes 1 % from all tweets. Although the *filter* method could gather all tracked tweets, this must not exceed the limit rate of 1 % of the whole Twitter stream. We note that only the paid *Firehose* method of Twitter API guarantees the delivery of 100 % of all tweets. However, previous work has shown that such obtained samples are close to the random samples over full Twitter stream [31]. We analyzed two tweet collections, restricted to English-language tweets, related to two crises:

1. *Sandy*: A dataset of tweets about Hurricane Sandy which was the most destructive hurricane in United States history with more than 230 deaths and 75 billion of damages. This dataset were collected from 29^{th} October 2012 to 31^{st} October 2012 using the 3 keywords: *"sandy"*, *"hurricane"* and *"storm"*.

2. *Ebola*: A dataset of tweets about West African Ebola virus epidemic which is the most widespread epidemic of Ebola virus disease. The epidemic began in Guinea in December 2013 and lasted for two years. World government agencies report more than $11,295$ deaths. The dataset were collected from July 29^{th} 2014 to August 28^{th} 2014 using *"ebola"* as a keyword.

The two datasets *Sandy* and *Ebola* have similar sizes with the number of tweets reaches 4,853,345 and 4,815,142, respectively.

The first key challenge was to select a sub-sample of informative tweets and filter out noisy ones. To tackle this issue, we processed the datasets using the methodology introduced in [8] for automatic data reliability detection. Practically, unreliable tweets are filtered using an automatic classifier. Based on regression logistic model, the automatic classifier maps tweets into two categories: *useful* and *useless*. The dataset processing is conducted into three steps:

- *Step 1: Building the training dataset.* A manual annotation task was assigned to 10 human assessors who were independently provided with *1)* instructions about the categorization task: a tweet is assessed as *useful* if it is related to the crisis and brings relevant information that helps to understand the tweet context or the situation. The tweet is *useless* otherwise; *2)* a set of manually annotated examples of each category (*useful* and *useless*), and *3)* a set of tweets from each crisis-related collection. The assessors were asked to choose a single category that best matched the content of the tweet. This task results in a training dataset including 1,800 labeled tweets and two dictionaries containing the 100 most frequent terms (resp. less frequent) from each category of manually annotated tweets (*useful* terms resp. *useless* terms). Using term frequencies extracted from annotated tweets would allow us to filter the most vs. less frequent topics embedded in the datasets.

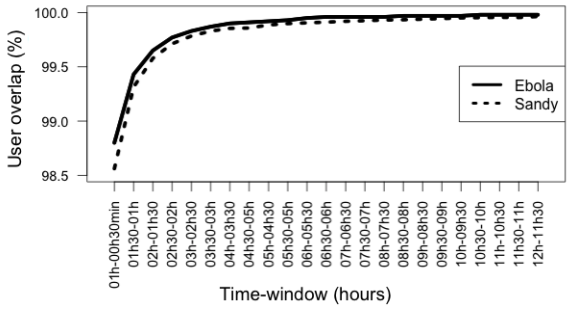
- *Step 2: Training the classification model.* We used the training dataset to learn the automatic classification model based on a logistic regression. To achieve this goal, we set up a feature-based representation of the manually annotated tweets based on 12 features that we classify into 3 categories: *1)* content features (e.g., number of hashtags, number of mentions), *2)* typographical features (e.g., number of punctuation characters, number of emoticons), *3)* vocabulary features (e.g., number of *useful* terms, number of *useless* terms). Since this feature is based on a statistical distribution of terms over the collection, as indicated above, we expect filtering useful tweets dealing with representative topics. This does not imply that useless tweets include only irrelevant tweets to the emergency situation. The performance of the training model was higher than 80 % for both datasets with a Mean Absolute Error (MAE) of 18 %.

- *Step 3: Filtering the datasets.* For the remaining analysis in this study, we only consider tweets classified as *useful* using the automatic classification model built in the previous step. Statistics about the resulting datasets are presented in Table 1.

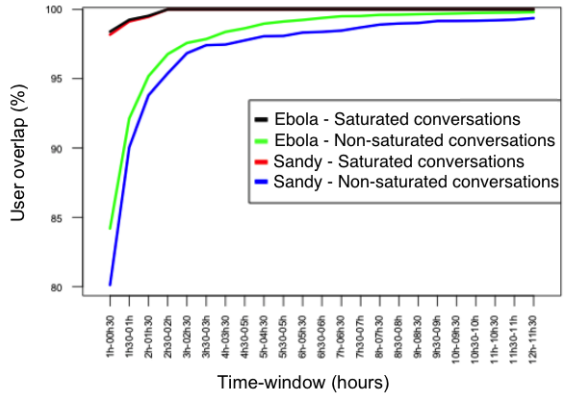
We can see from Table 1 that both datasets contain more than 40 % of retweets, 64 % of mentions, and between 28 % and 46 % of shared URL. These statistics clearly show the engagement of the users involved in these datasets to explicitly collaborate through conversation built as detailed below.

Dataset	Sandy	Ebola
Tweets	2,119,854	2,872,890
Microbloggers	1,258,473	750,829
Retweets	963,631	1,157,826
Mentions	1,473,498	1,826,059
Reply	63,596	69,773
URLs	596,393	1,309,919
Pictures	107,263	310,581

Table 1: Descriptive statistics of each crisis-related dataset.



(a) Overlap ratio over all conversations



(b) Overlap ratio over saturated and non-saturated conversations

Figure 1: Tuning the temporal parameter.

3.3 Conversation Datasets

We assume that explicit collaboration in online social networks is channeled by a certain type of interactions. With respect to the definitions presented in section 3, we aim at describing and differentiating interactional patterns which we deem to be typically collaborative. According to Assumption 1, we built the set of conversations, based on the vocabulary of Twitter user interactions, namely *replying*, *mentioning* or *retweeting*, all mediated by the use of the @ symbol which conveys a collaboration manifestation [11]. Practically, the @ symbol is followed by the user handle (username) and thereby defines a link from tweet to tweet and, further, from user to user.

In order to characterize such conversations, we applied the algorithm proposed by Cogan et al. [9] on the cleaned data sets including only useful teweets. The algorithm works in two steps: starting from a given tweet, it first goes upstream by recursively discovering tweets in an ascending manner until it finds the root tweet. It then goes downstream, in a descending manner, to explore the subset of tweets related

Dataset	Sandy - 2 hours		Ebola - 2 hours	
# tweets	758,887	(79.83%)	878,171	(79.27%)
# user	1,020,213	(84.24%)	1,102,895	(83.17%)
# *retweet*	702,227	(78.75%)	825,642	(78.77%)
# *reply*	56,682	(96.28%)	52,624	(88.06%)
# *mention*	90,370	(83.51%)	157,457	(80.23%)
# conversation	240,991	(100.00%)	196,005	(100.00%)
Average number of tweets per conversation	3.15	(79.83%)	4.48	(79.27%)
Average number of users per conversation	4.23	(84.24%)	5.62	(83.17%)
Average number of *retweets* per conversation	2.91	(78.75%)	4.21	(78.77%)
Average number of *reply* per conversation	0.24	(96.28%)	0.27	(88.06%)
Average number of *mention* per conversation	0.37	(83.51%)	0.80	(80.23%)
Average length of the conversation	1.10	(99.68%)	1.23	(98.83%)

Table 2: Descriptive statistics for the 2 hours conversations for Sandy and Ebola datasets. Percentage (%) represents the divergence index between temporally truncated trees and complete trees, using 100.00 as a basis for complete trees. Most index are close to 100.00.

to this root tweet. Put differently, it reconstructs the whole tree of conversation in which the original tweet is embedded. On the Sandy and Ebola datasets, we respectively gather $240,991$ and $196,005$ conversation trees.

According to Assumption 2, a conversation may diverge progressively towards subgroups of users and topics, long after it started, plausibly far from the original collaborative search and sharing task goal. Indeed, in graph-theoretical terms, conversation trees are downstream connected components which can grow very large in practice. The chain of interactions may go deep and it is reasonable to put a boundary on long-lasting discussions: explicit collaboration should a priori correspond to compact interaction patterns, both structurally, temporally, and topically. A conversation may indeed diverge progressively towards subgroups of users, topics, long after it started plausibly far from the original collaboration goal.

To deal with this issue, we introduce a simple criterion to limit the conversation tree exploration: we require downstream tweets to be sufficiently close in time to the root tweet, whereby no more than T minutes should separate the last tweet from the root. This approach relies on the notion that conversations reach a point of saturation hereafter the explicit collaboration group is relatively stable. To check this, we examine the social content of trees as a function of time from the root tweet. More precisely, we look at the temporal evolution of participant lists when choosing a longer exploration period. We compute the average ratio of overlap between user lists, for a same conversation collected until t from the root seed and $t + 30$ minutes.

Figure 1 shows the distribution of the collaborative group overlaps over time for both Sandy and Ebola datasets. The x-axis represents the two successive temporal constraints (t and $t + 30$), while the y-axis corresponds to this *overlap ratio*. We notice that fixing T at 120 minutes yields an average overlap ratio of more than 99 %, for both datasets. We thus distinguish "*saturated*" from "*non-saturated*" conversations by considering trees which exhibit an overlap ratio of 100 % after 120 minutes (or not). Even for non-saturated conversations, we observe an average overlap ratio around 95 %. When comparing temporally-truncated trees with complete trees (see Table 2), we still observe a strong similarity in terms of structural features which further justifies the choice of this constraint. Truncated trees constitute the basic structural pattern of user groups engaged in explicit collaboration bounded by a relatively tight time constraint. We can now analyze both their social and semantic configurations.

4. RESULTS

4.1 Characterization of the Collaborative Groups

Our objectives here are twofold: *1)* identifying the structural patterns of the users' collaboration networks and *2)* identifying the topics of the conversation threads that might make sense of the crisis situation for the users engaged in conversations.

4.1.1 Structure of the collaborative group networks

To have a picture of the collaboration structure, we derive the interaction subgraphs from the conversation trees. Nodes of the subgraph are conversation tree participants, while links correspond to at least one interaction between two users. We are particularly interested in the shape of these subgraphs and their distribution as done in [27, 15] for the study of diffusion cascade patterns in blog post networks. However, we are dealing here with conversation patterns, i.e., social subgraphs based on mentions, retweets and replies, in order to eventually describe distinct conversation roles and configurations [12].

To start with, we introduce a nomenclature defining a subgraph by a pair of numbers (x, y) where x represents the number of users and y the number of links. An analysis of this simple notation already makes it possible to describe and discriminate a wide range of different subgraphs. Table 3 gathers the most frequent interaction subgraphs and their graphical representations. Interestingly, this taxonomy is roughly identical for both Ebola and Sandy datasets, both in terms of rank and order of magnitude of the respective patterns. There are essentially two main types of subgraphs which appear to be meaningful in terms of potential collaboration configurations, involving between 2 and 7 users:

1. *Star-shaped networks* (patterns $(2, 1)$, $(3, 2)$, $(4, 3)$, $(6, 5)$, etc.). These subgraphs feature the prominence of a central person, and subsequent discussants which are all linked to her. Qualitatively, they correspond to information relaying subgroups where peripheral indi-

Setting	#Sandy	#Ebola	Pattern	Setting	#Sandy	#Ebola	Pattern
2;1	157,687	96,573		3;2	36,929	35,694	
4;3	12,124	11,639		5;4	5,568	6,058	
4;4	2,767	4,342		6;5	3,394	3,855	
7;6	2,177	2,862		8;7	1,528	2,434	
5;6	1,446	2,322		3;3	750	2,181	

Table 3: Most frequent collaboration patterns in Sandy and Ebola datasets, for saturated conversations.

viduals all cite or retransmit the content published by the central person.

2. *Flatter networks* (patterns $(4,4)$, $(3,3)$ or $(5,6)$, i.e., which respectively look like a square, a triangle, or a square and a triangle). These subgraphs indicate a more horizontal, collective discussion structure where more users interact with each other in a relatively decentralized manner, that is, without the existence of a central user.

Overall, we can see that collaborative groups are quite small without necessarily involving central users.

4.1.2 *Content of the intra-group interactions*

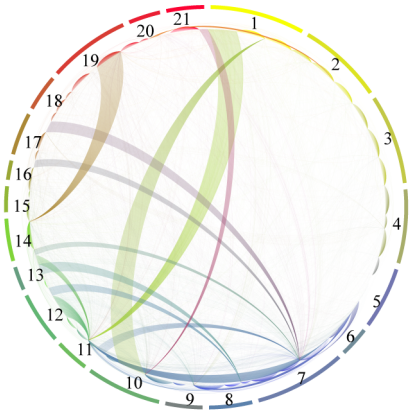
With respect to our second goal related to the topical analysis of the conversation streams, we applied the Latent Dirichlet Allocation (LDA) model [4] to the meta-documents built from the conversations and then tuned the optimal number of parameters using the perplexity measure [4]. We reached a minimal perplexity at 16 topics for the Ebola collection and 21 topics for the Sandy collection. Three assessors made a manual unsupervised annotation of the topics automatically extracted with the LDA model, to define topic labels. In case of disagreement on topic labels, a consensus has been reached between the three assessors. As shown by the labels listed in Table 4, the topics extracted from both collections are mostly related to crisis management except some of them (e.g., Obama and the Benghazi attack) which could be due to classification errors (18 %). We outline that these topics are quite different from those identified in similar emergency situations [22] extracted from the tweet streams, regardless of the groups' conversations. This observation is expected since topics extracted from conversations are likely to be more focused.

In order to relate the group structure to the underlying task topic, we associated each conversation with a topic through the maximal probability assignment criteria in the distribution conversation-topic resulting from the LDA model. Table 5 shows statistics about conversations by topic. We can notice that the most represented topics according to conversation numbers (#Conv.) related to crisis management are *1)* prayers, negative thoughts and thanks, for the Sandy collection, and prevention, victims/quarantine, and *2)* actions/thoughts to people, for the Ebola collection.

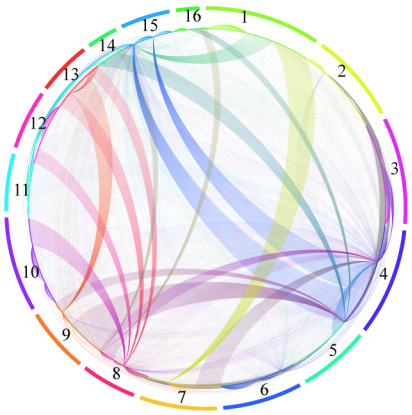
Dataset	Topics
Sandy	(1) State of New York City; (2) Negative thoughts; (3) Donations/aids; (4) Thanks; (5) Explanations; (6) Water/Flood; (7) Insults; (8) Photos/Videos; (9) Dead persons/Deaths; (10) After Sandy; (11) Damages; (12) Missing people; (13) Prayers; (14) Obama and the Benghazi attack; (15) Weather alerts and nuclear alerts; (16) Humor; (17) Fear/Terror; (18) Financial impact; (19) Report/Inventory of fixtures; (20) Communication tools; (21) Information via the media
Ebola	(1) Prevention; (2) Actions/Thoughts to people; (3) Official reports; (4) Personal thoughts; (5) Dead persons/Deaths; (6) Worldwide urgency; (7) Exile; (8) Propagation; (9) Clinical tests; (10) Drug/Vaccine research; (11) Treatments; (12) First case in the US; (13) Disease/Fear in the US; (14) Victims and quarantine; (15) Action plan in Africa; (16) Propagation control

Table 4: Extracted conversation topics from Sandy and Ebola datasets.

To characterize the network of the collaborative groups related to each topic, we computed the modularity measures. This metric measures the relationship density between and within users' modules obtained by a classification algorithm. We found a high average modularity value that reaches 0.96 for Sandy dataset and 0.88 in the case of Ebola dataset. These modularity values reflect a higher density between the users of collaborative groups but -in contrast- sparse connections between users of different collaborative groups. The computation of the ratio between the number of identified modules and the number of conversations for each topic revealed a high ratio (0.8) for Sandy and a low ratio for Ebola (0.3). The results obtained on the Sandy dataset suggests that conversation networks are highly disconnected while those obtained on the Ebola dataset sug-



(a) Sandy dataset



(b) Ebola dataset

Figure 2: Collaboration networks

	Sandy				Ebola		
	#Conv.	#Users	#Tweets		#Conv.	#Users	#Tweets
1	12,325	49,979	36,618	1	42,811	238,022	188 987
2	17,857	74,853	55,493	2	13,282	71,808	56,656
3	12,573	55,557	41,938	3	10,640	56,377	44,240
4	13,779	56,475	41,509	4	9,846	54,375	43,153
5	9,743	41,583	30,974	5	10,103	54,849	43,166
6	10,365	41,108	29,887	6	9,041	51,871	41,478
7	38,781	163,959	121,852	7	12,950	75,458	60,597
8	10,152	41,505	30,482	8	8,719	49,632	39,629
9	8,297	34,678	25,642	9	7,081	38,012	29,932
10	8,928	36,354	26,703	10	11,294	64,080	51,103
11	6,842	28,256	20,818	11	5,762	32,750	26,151
12	9,642	42,084	31,557	12	11,383	70,132	57,151
13	23,329	102,708	77,465	13	6,925	41,058	33,032
14	9,477	41,560	31,273	14	20,594	114,007	90,347
15	11,672	51,118	38,489	15	8,374	49,266	39,660
16	4,633	19,199	14,234	16	7,202	41,200	32,888
17	5,461	24,195	18,339				
18	8,278	35,337	26,313				
19	5,897	23,053	16,666				
20	6,817	29,855	22,496				
21	6,164	26,797	20,138				

Table 5: Statistics of the collaborative groups sharing similar interests.

gest that a high ratio of conversations are subsumed. The difference in these observations could be explained by the differences in the life-spans of the datasets as explored in the next section. It seems that long-time social interactions lead to create user-to-user connections and bridge the gap between some of the collaborative groups.

In summary, these results outline that many collaborative groups deal with the same topic, but these groups are not totally connected as they form distinct ones, particularly within a short-term period. This may be explained by the fact that the conversations were held synchronously or asynchronously by users with similar interests who fail to connect with each other. To get a deeper understanding of this observation, we focus below on the connectivity between the collaborative groups.

4.2 Examination of the Connectivity between the Collaborative Groups

We believe that an interesting analysis concerns the exploration of the global collaboration network in order to highlight whether and how groups of users with similar interests (or not) are likely to explicitly collaborate with each other. Therefore, it seems interesting to identify relationships (reply, retweet, and mention links) between conversations, whether they share similar interests or not.

Respectively for Sandy and Ebola datasets, Figures 2a and 2b illustrate the collaboration networks inferred from users' social interactions considering their LDA topic-based assignation (see Section 4.1). These networks include the whole overlapping conversations, from the user belonging perspective, dealing with the same topic. Since the datasets include a large amount of users and social interactions, and in order to favor a better readability, we have represented these collaboration networks by filtering the three most populated conversations within each LDA topic. In each figure, collaboration topics are scattered around the perimeter of the circle and two types of relationships are characterized:

1. The "*intra-topic*" relationships between users engaged in conversations addressing the same topic, illustrated by arcs around and external to the perimeter. These relationships as illustrated in these figures are not particularly dense. We remember that Figure 2.a and Figure 2.b illustrate only the top three conversations of each topic which explain the low visible density. With all conversations considered, the number of average of "intra-topic" relationships exceeds 47, 700 relations for Sandy data and 36, 300 relations of Ebola dataset. Moreover, the degree of density varies across topics. These results added to structural analysis of the collaboration structure (see 4.1) corroborates the lack of social connectivity between users engaged in similar search/sharing tasks.

2. The *"inter-topic"* relationships between users engaged in distinct conversations dealing with distinct topics, symbolized by arcs inside the circle, are more prevalent for Sandy dataset whereas Ebola is characterized by a greater number of "inter-topic" relationships.

Considering the fact that Figures 2a and 2b show a subrepresentation of the overall collaboration networks, we first

	Sub-Dataset 1	Sub-Dataset 2	Sub-Dataset 3
Period	07/29/2014 07:23 am - 07/31/2014 9:59 pm	08/07/2014 9:39 am - 08/08/2014 5:54 pm	08/14/2014 10:00 pm - 08/17/2014 9:59 pm
#Tweets	394,399	311,340	326,985
#Users	165,764	137,398	123,260
Cohesion value	1.418	1.309	1.958

Table 6: Analysis of the impact of the time-window on the cohesion metric.

Sandy	Ebola
(7) Exile - (13) Disease/Fear in the US	(1) Prevention - (14) Victims and quarantine
(7) Exile - (1) State of New York City	(1) Prevention - (2) Actions/Thoughts to people
(7) Exile - (2) Negative thoughts	(1) Prevention - (7) Exile
(7) Exile - (3) Donations/aids	(1) Prevention - (12) First case in the US
(2) Negative thoughts - (13) Disease/Fear in the US	(2) Actions/Thoughts to people - (14) Victims and quarantine

Table 7: Top pairwise topics assigned to users and inferred from their conversations.

judge the generalizability of our observations by the use of the cohesion metric [5] estimated over the collaboration network obtained with the whole datasets characterized by topical clusters associated to the LDA topics. This metric estimates the ratio between the strength of "*intra-topic*" user relationships (within topical cluster) and the strength of the "*inter-topic*" user relationships (between topical clusters), where a value higher than 1 highlights the preponderance of "*intra-topic*" relationships. We obtained a cohesion metric value equals to 1.43 for Sandy and 0.63 for Ebola, suggesting that the previous observed statements on the network sub-representation could be generalized to the overall collaboration network.

The differences between the Ebola and Sandy collaboration networks could be explained with respect to the duration of the social activity traced to the beginning of the crisis, at least starting from the oldest timestamp of the tweets in the crisis-related datasets. Indeed, both datasets, although characterized by similar collaboration patterns in individual conversations (see Table 3), have contrasted results in terms of cohesion metric values. Due to the collect time-periods (3 days for the Sandy dataset and 1 month for the Ebola one), we could infer that collaboration relationships of those datasets are respectively performed in short and long-term. Accordingly, the cohesion metric suggests that short-term interactions between users are more likely to focus on few topics ("*intra-topic*" relationships), in contrast to long-term interactions that seem to provide a wider range of topics and favor "*inter-topic*" social interactions. This result is consistent with previous findings [17] which indicate that the nature and intensity of communication between users change over-time according to the crisis phases. We expect that the latter has a significant effect on the degree of diversity of the sub-topics since an increasing number of users are involved during the evolving sense-making process.

To deepen our understanding of the differences between both datasets according to the time period parameter, we split the Ebola dataset into several sub-datasets in which the collect period is relatively equivalent to the one of the Sandy dataset. We randomly selected 3 sub-datasets in order to identify conversations, extract topics and finally built the collaboration network. We present in Table 6 the characteristics of these 3 sub-datasets and the obtained cohesion metrics. We notice that the cohesion metric values are higher than 1 (respectively equals to 1.418, 1.309 and 1.958 for the 3 sub-datasets), reinforcing our intuition that short-term interactions favor collaboration around few and more focused topics, with less interest sharing with the other users. This observation leads us to argue that large-scale collaboration is timely influenced by the emergence (or not) of users playing roles of intermediaries between distinct groups of users.

Moreover, we highlight that users are generally implied in different conversations, which might favor the diversity of the topic assignment at the user level. Indeed, for the Ebola dataset, a user is assigned on average to 7 topics while this value is lower (equals to 4) in the Sandy dataset. Therefore, it seems reasonable to assume that a user engaged in multiple conversations with different topics is more likely to create "*inter-topic*" relationships than a user engaged in few or a unique conversation. For instance, the most identified pairwise topics assigned to users within each dataset are represented in Table 7. We notice that for both datasets, one topic remains predominant in pairwise associations, respectively topic 7 and 1 for Sandy and Ebola, highlighting its exploratory aspect. This is mostly identifiable for Ebola since prevention might be carried according to several dimensions. This topic connectedness highlights the fact that some topics (likely exploratory topics) include a wide range of subtopics (connected topics), leading to consider sub-topics as micro-tasks built naturally over users' conversations.

5. IMPLICATIONS FOR SOCIAL MEDIA COLLABORATION RESEARCH

We believe that our findings have important implications relevant for the research in social web collaboration, including:

• *Recommendation of collaborators.* Our findings indicate that the social graph of collaborative groups of users engaged in similar or shared search and sharing tasks is a set of weakly connected (or disconnected) small-sized sub-graphs. People are likely to be connected with a small audience while, being involved in an open social web space, they believe that they are connected to the crowd. One relevant research opportunity would be to enhance the collaboration mediation between users by designing information seeking algorithms able to automatically

enhance user's seeking tweets with mentions leading to collaborator recommendation. This would allow creating social collaborative ties between users in order to overpass the small-sized collaboration network restricted to the user's neighborhood guided by his mentions and followers. Such algorithms would *1)* tackle the issue mentioned in [23] related to unanswered seeking tweets and *2)* constitute a valuable support for users who explicitly asked the questions to the crowd using the hashtag rather than their followers' network, as revealed in [17]. In addition, algorithms could be created to identify long-term group members who reinforce collaboration by providing advice or help through the identification of appropriate contributors for solving topically-related search and sharing tasks. The presence of a strong social support on the network is likely to encourage low-activity users to be better engaged towards a larger audience and benefit from it.

• *Enhancement of social awareness.* The analysis of collaborative groups connectivity demonstrated that active users are involved in many conversations with distinct topics, either sub-topics of the root search/sharing task topic or even other tasks dealing with other topics. Thus, another interesting implication for future research would be to design algorithms that enhance the information seeking process by detecting complex information needs and tasks. These would then be rooted to users engaged in different collaborative groups of users through conversations gathering shared or complementary topical facets of the original information need and task. Such users, viewed here as valuable intermediaries, could better transfer the need or task through the social network and increase more quickly the situational awareness rate over the whole social network. In turn, this would enhance the self-engagement of users, which is particularly desired in emergency situations. More generally, algorithmic mediation on the social web would create social ties between users or turn social ties into collaborative ties, tackling the problem of the so-called *"digital desert"* [3].

6. STUDY LIMITATIONS AND FUTURE WORK

The study faces some limitations. First, we acknowledge that the generalization of our findings requires further work. One limitation is the non-diversity and size of samples used in the study. For future research, we plan to perform our study across large-scale collections of User Generated Content (e.g., forums) within different application domains or search and sharing intents–beyond crisis-management–(e.g., health concerns). This would provide additional insights into how social media-based collaboration occurs.

We also note the lack of data about users. That is, based on our findings, we are not able to characterize group patterns according to the social neighborhood of the group members. This requires tracking users' personal profiles as well as those of their followers and those of users they mention or retweet. We intend to collect these data in the future to evaluate the ratio of strangers and their adequacy, compared to social neighbors, to solve the search topic. User-centric data would likely help us better explain the absence of explicit collaboration from a user perspective.

We finally point out that while previous works have shown that seeking and sharing information are the two basic forms of online explicit collaboration using social networks [10, 29], we analyzed them in this study without distinction. The latter would be possible by building distinct conversation trees considering tweet-question seeds vs. non tweet-question seeds. Hence, we were not able to determine the differences that these two forms of collaboration might exhibit in both structural patterns and topics. We plan to overcome this limitation in the near future and explore the statistical differences between the corresponding social graphs.

7. CONCLUDING REMARKS

We studied explicit collaborative information search/sharing practices supported by social media. We used data from two crisis-related Twitter datasets. We first generated conversation datasets built upon a target collaboration concept and then analyzed their social structure and content. The results of the analyses reported here particularly highlight that *1)* collaboration is generally limited to small-sized flat networks, with or without a central user, based on user's explicit mentions, replies, and retweets, *2)* users are often engaged in distinct conversations involving different users, members of disconnected or weakly connected groups, and dealing with both distinct and shared topics, and that *3)* the time factor impacts the structure of the collaboration network; more particularly, we show the existence of a robust threshold in the time-span of conversation trees.

Based on our findings, we provided relevant research opportunities that would enable the emergence of a new generation of social collaborative information sharing and search systems.

8. ACKNOWLEDGMENTS.

This research was supported by the French CNRS PEPS research program under grant agreement EXPAC (CNRS/PEPS 2014-2015).

9. REFERENCES

[1] A. Acar and Y. Muraki. Twitter for crisis communication: Lessons learned from japan tsunami disaster. *Int. J. Web Based Communities*, 7(3):392–402, 2011.

[2] I. Badache and M. Boughanem. Harnessing social signals to enhance a search. In *WI*, volume 1, pages 303–309, Aug 2014.

[3] R. Baeza-Yates and D. Saez-Trumper. Wisdom of the crowd or wisdom of a few? an analysis of users' content generation. In *Hypertext and Social Media*, pages 69–74. ACM, 2015.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[5] R. D. Bock and S. Z. Husain. An adaptation of holzinger's b-coefficients for the analysis of sociometric data. *Sociometry*, 13(2):pp. 146–153, 1950.

[6] F. Cacheda, V. Carneiro, D. Fernández, and V. Formoso. Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. *ACM Trans. Web*, 5(1):2:1–2:33, Feb. 2011.

[7] D. Carmel, N. Zwerdling, I. Guy, S. Ofek-Koifman, N. Har'el, I. Ronen, E. Uziel, S. Yogev, and

S. Chernov. Personalized social search based on the user's social network. In *CIKM*, pages 1227–1236. ACM, 2009.

[8] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *WWW*, pages 675–684. ACM, 2011.

[9] P. Cogan, M. Andrews, M. Bradonjic, W. S. Kennedy, A. Sala, and G. Tucci. Reconstruction and analysis of twitter conversation graphs. In *HotSocial Workshop*, pages 25–31. ACM, 2012.

[10] M. De Choudhury, M. R. Morris, and R. W. White. Seeking and sharing health information online: Comparing search engines and social media. In *CHI*. ACM, 2014.

[11] K. Ehrlich and N. S. Shami. Microblogging inside and outside the workplace. In *ICWSM*. The AAAI Press, 2010.

[12] I. Eleta and J. Golbeck. Multilingual use of twitter: Social networks at the language frontier. *Computers in Human Behavior*, 41:424–432, 2014.

[13] B. M. Evans and E. H. Chi. Towards a model of understanding social search. In *CSCW*, pages 485–494. ACM, 2008.

[14] J. Foster. Collaborative information seeking and retrieval. *Annual Rev. Info. Sci & Technol.*, 40(1), Dec. 2006.

[15] L. Franco and H. Kawai. News detection in the blogosphere: Two approaches based on structure and content analysis. In *ICWSM*, 2010.

[16] J. P. Gene Golovinsky and M. Back. A taxonomy of collaboration in online seeking. In *arxiv.org/pdf/0908.0704*, CoRR abs/0908.0704, 2009.

[17] T. Heverin and L. Zach. Use of microblogging for collective sense-making during violent crises: A study of three campus shootings. *JASIST*, 63(1):34–47, 2012.

[18] S. R. Hiltz, P. Diaz, and G. Mark. Introduction: Social media and collaborative systems for crisis management. *ACM Trans. Comput.-Hum. Interact.*, 18(4):18:1–18:6, 2011.

[19] T. Holderness and E. Turpin. *PetaJakarta.org: Assessing the Role of Social Media for Civic Co-Management During Monsoon Flooding in Jakarta, Indonesia*. SMART Infrastructure Facility, University of Wollongong, 2015.

[20] D. Horowitz and S. D. Kamvar. The anatomy of a large-scale social search engine. In *WWW*, pages 431–440. ACM, 2010.

[21] M. Imran, C. Castillo, F. Diaz, and S. Vieweg. Processing social media messages in mass emergency: A survey. *ACM Comput. Surv.*, 47(4):67:1–67:38, June 2015.

[22] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier. Practical extraction of disaster-relevant information from social media. In *WWW*, pages 1021–1024, 2013.

[23] J.-W. Jeong, M. R. Morris, J. Teevan, and D. Liebling. A crowd-powered socially embedded search engine. In *ICWSM*. AAAI, 2013.

[24] B. Karweg, C. Huetter, and K. Böhm. Evolving social search based on bookmarks and status messages from social networks. In *CIKM*, pages 1825–1834. ACM, 2011.

[25] A. Kavanaugh, S. D. Sheetz, F. Quek, and B. J. Kim. Cell phone use with social ties during crises: The case of the virginia tech tragedy. *Using Social and Information Technologies for Disaster and Crisis Management*, (84), 2013.

[26] V. Lampos, T. De Bie, and N. Cristianini. Flu detector: Tracking epidemics on twitter. In *ECML PKDD*, pages 599–602. Springer-Verlag, 2010.

[27] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. 2007.

[28] J. T. Meredith Ringel Morris and K. Panovitch. What do people ask their social networks, and why? a survey study of status messages q&a behaviour. In *CHI*, pages 1739–1748. ACM, 2010.

[29] M. R. Morris. Exploring the complementary roles of social networks and search engines. In *ICWSM*, 2011.

[30] M. R. Morris. Collaborative search revisited. In *CSCW*, pages 1181–1192. ACM, 2013.

[31] F. Morstatter, J. Pfeffer, and H. Liu. When is it biased?: assessing the representativeness of twitter's streaming api. In *WWW*, 2014.

[32] L. Palen and K. Starbird. Working and sustaining the virtual disaster desk. In *CSCW*, 2013.

[33] J. Pickens. Social and collaborative information seeking: Panel. In *CIKM*, pages 2647–2648, New York, NY, USA, 2011. ACM.

[34] S. Poltrock, J. Grudin, S. Dumais, R. Fidel, H. Bruce, and A. M. Pejtersen. Information seeking and sharing in design teams. In *SIGGROUP*, pages 239–247, 2003.

[35] A. Sadilek, H. A. Kautz, and V. Silenzio. Modeling spread of disease from social interactions. In *ICWSM*, 2012.

[36] C. Shah. *Collaborative Information Seeking - The Art and Science of Making the Whole Greater than the Sum of All*, volume 34 of *The information retrieval series*. Springer, 2012.

[37] L. Soulier, C. Shah, and L. Tamine. User-driven system-mediated collaborative information retrieval. In *SIGIR*, pages 485–494, 2014.

[38] L. Soulier, L. Tamine, and W. Bahsoun. On domain expertise-based roles in collaborative information retrieval. *IP&M*, 50(5):752 – 774, 2014.

[39] J. Surowiecki. The Wisdow of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations. *Random House*, pages 78–85, 2004.

[40] J. C. Tang, M. Cebrian, N. A. Giacobe, H.-W. Kim, T. Kim, and D. B. Wickert. Reflecting on the DARPA Red Balloon Challenge. *Commun. ACM*, 54(4):78–85, 2011.

[41] S. Vieweg, A. L. Hugues, K. Starbird, and L. Paeln. Microblogging during two natural hazards events: What twitter may contribute to situational awarness. In *CHI*, 2010.

[42] C. Wendling, J. Radisch, and S. Jacobzone. The use of social media in risk and crisis communication. *OECD Working Papers on Public Governance*, (24), 2013.