Answering Twitter Questions: a Model for Recommending Answerers through Social Collaboration
Laure Soulier, Lynda Tamine, Gia-Hung Nguyen

To cite this version:

HAL Id: hal-01353587
https://hal.sorbonne-universite.fr/hal-01353587
Submitted on 12 Aug 2016

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.
Answering Twitter Questions: a Model for Recommending Answerers through Social Collaboration

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

Lynda Tamine
Université de Toulouse,
UPS - IRIT
118 route de Narbonne
31062 Toulouse, France
tamine@irit.fr

Gia-Hung Nguyen
Université de Toulouse,
UPS - IRIT
118 route de Narbonne
31062 Toulouse, France
gia-hung.nguyen@irit.fr

ABSTRACT

In this paper, we specifically consider the challenging task of solving a question posted on Twitter. The latter generally remains unanswered and most of the replies, if any, are only from members of the questioner’s neighborhood. As outlined in previous work related to community Q&A, we believe that question-answering is a collaborative process and that the relevant answer to a question post is an aggregation of answer nuggets posted by a group of relevant users. Thus, the problem of identifying the relevant answer turns into the problem of identifying the right group of users who would provide useful answers and would possibly be willing to collaborate together in the long-term. Accordingly, we present a novel method, called CRAQ, that is built on the collaboration paradigm and formulated as a group entropy optimization problem. To optimize the quality of the group, an information gain measure is used to select the most likely “informative” users according to topical and collaboration likelihood predictive features. Crowd-based experiments performed on two crisis-related Twitter datasets demonstrate the effectiveness of our collaborative-based answering approach.

Keywords
Social information retrieval, Collaborative group recommendation, Social Network Question-Answering

1. INTRODUCTION

A recent Pew Internet survey\(^1\) published in August 2015 indicates that Facebook and Twitter are the most prominent social media services, with more than 72% and 23% of US adults utilizing the services, respectively. Although social platforms were designed to create social connections, they have emerged as tractable spaces for information seeking through the posing of questions [20]. A survey on question-asking practices on Twitter and Facebook reveals that over 50% of users ask questions on social networks [40]. The main reason users prefer a social platform to a search engine is that they trust their social network’s replies and specifically seek subjective replies (e.g., opinions/advice) rather than objective information provided by search engines.

However, there are several issues related to asking questions on social media. One issue is that the majority of questions do not receive a reply, while a minority receives a high number of responses [21, 32]. Another issue is that even though questioners use manual question-oriented hashtags (e.g., #lazyweb, #twoogle) to bypass the local neighborhood [19], the answers are mostly provided by members of the immediate follower network, characterizing behavior patterns known as friendsourcing [20]. These issues give rise to other drawbacks. First, users are uncomfortable with posing private questions (e.g., religious or political) to their social neighbors [29]. Second, recent research has highlighted that questioners perceive that friendsourcing has a social cost (e.g., spent time and deployed effort) that weakens the potential of social question-answering [20, 30].

Hence, an effective recommendation for skilled answerers responding to questions posted online for a wide audience is highly desirable. In this perspective, a common approach early developed in community Q&A services [24, 22] and in other popular social networks, such as Facebook and Twitter [17, 25], consists in routing the questions to a list of the top-k appropriate users. The users’ appropriateness is generally estimated using a set of features from the questions, the users themselves and their relations with the questioner. However, collaboration has been acknowledged as a valuable method for completing information search tasks within [28, 9] (e.g., the DARPA challenge [39]) and outside of social media spaces [12]. Therefore, another approach that we believe to be valuable, is to consider the question-answering task as a collaborative process involving a group of socially authoritative users with complementary skills. Such users would allow for the gathering of diverse pieces of information related to the question, thereby reinforcing the likelihood that their answers are relevant as a whole. Accordingly, we address the problem of tweet question solving by turning it into the problem of collaborative group building which is the core contribution of this paper. More specifically, based on the information gain theory, our approach favors the selection of a collaborative group of answerers rather than the ranking of answerers. Our long-term goal is to favor explicit social collaboration between the group members including the questioner.

\(^1\)http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/
Based on this driving idea, we propose the CRAQ- Collaborator Recommendation method for Answering Twitter Questions. Given a posted question q and topically relevant tweets T posted by an initial group U of users, the general schema supporting the CRAQ relies on an evolving process of group building that discards users who supply the lowest information gain at each step, leading to the maximization of the group information entropy [33]. To summarize, this paper makes the following contributions:

- First, we introduce a novel algorithm recommending a collaborative group of users to tweeted questions so as a cohesive answer could be inferred from the aggregation of their topically relevant tweets. To the best of our knowledge, this is the first effort to recommend a social collaborative group of users aiming at maximizing the overall relevance of answers to a tweeted question.
- Second, we build a predictive model to characterize the collaboration likelihood between pairwise users who are collectively able to solve target tweeted questions. The model is built on both the user’s authority and the complementarity of the content of their tweets.
- Third, we perform an empirical crowdsourced-based evaluation which shows that collaborative-based tweet question answering is effective.

In what follows, Section 2 reviews the relevant prior work. Section 3 presents the problem and provides an overview of the CRAQ. Section 4 details the predictive model of collaboration and the group recommendation algorithm. Section 5 details the experimental setup, and the results are presented in Section 6. Last, we conclude in Section 7.

2. RELATED WORK
Social Network Question Answering (SNQ&A).

Unlike in community Q&A [24, 22, 9], users who post questions on social network sites are engaged in information seeking with a wide range of users from their social cluster or even strangers. Numerous research studies have investigated the characteristics of questions addressed to social networks, including the topics and the motivations behind the questions [29, 32, 21]. Through a survey of Facebook and Twitter users, Morris et al. [29] showed that most users post questions to obtain subjective (opinion-based and engaged) answers related to a wide range of interests, including technology, entertainment, home, and family. Considering the benefits of engaging in social collaboration while performing a search [17, 28, 15], Horowitz et al. [17] designed the Aard-vark social search engine built on the “village paradigm". The authors’ approach consists in routing the question to different people using a retrieval task aiming at identifying the appropriate person to answer the question. Formally, the authors proposed a probabilistic model based on (1) users’ expertise, (2) the connectedness of the recommended user and the questioner, and (3) the availability of the candidate answerer. Similarly, Hecht et al. [15] presented a socially embedded search engine, called Search Buddies and collocated with Facebook. This system includes two main functionalities that are proposed to answer questions: recommending relevant posts or messages that are likely to answer a question (Investigator API) and connecting to people who may have the answer (Butterfly API). One limitation of these two previous studies is that they tracked information and users within the questioners’ neighborhood. This method might be restrictive and hinder the likelihood of obtaining replies [21, 32]. Unlikely, one recent work [25] which is most related to this paper attempt to automatically identify appropriate users, even strangers, to answer tweeted questions. The users’ appropriateness was estimated based on their willingness to answer the question and readiness to answer the question shortly after it was submitted. Statistical models were used to predict the probability that a user would provide an answer to the question. Second, based on the likelihood that a user answers the question, a ranked list of users is retrieved using a classification algorithm identifying users as responders or non-responders. Experiments based on three datasets collected via a human operator who presented questions to target users on Twitter showed that the recommendation algorithm improves the response rate compared with random users (by 59%) and a simple binary classification of target users (by 8%). In contrast to these authors’ work, we aim to recommend a collaborative group of users viewed here as authors of diverse and complementary pieces of answers with a maximal expected relevance to a given question. To the best of our knowledge, it is the first attempt in the context of SNQ&A for providing a collaborative-based answer to tweet questions. Beyond the goal of providing the relevant answer to a tweet question, our method can be applied to recommend users and then favor long-term collaboration among questioners and answerers.

Group selection.

Group selection [2, 13] is one of the key underlying issues addressed in a long-standing and multidisciplinary study on collaboration [3]. Regardless of the contexts in which it is embedded, explicit (implicit) collaboration typically refers to individuals’ action of intentionally (unintentionally) working together to complete a shared task. For instance, Augustin-Blas et al. [2] proposed a genetic algorithm-based model for group selection. The objective function of the algorithm consists of maximizing the shared knowledge resources within the selected candidate teams. This approach has been shown to be effective in solving the issue of teaching group formation. The study conducted by González-Ibáñez et al. [13] focused on estimating the costs and benefits of pseudo-collaboration between users with similar information search tasks. In pseudo-collaboration, users performing a search task are given recommendations based on the past results obtained by other users who performed similar tasks. An experimental study revealed that the efficiency of group selection based on pseudo-collaboration was greater than that the one of the group selection including users who intentionally shared a search task through explicit collaboration. More recent studies specifically addressed group selection within social media platforms [8, 6]. Castilho et al. [8] investigated whether the social behavior of candidate collaborators is an important factor in their choices when forming a collaborative task team. The authors performed an analysis using data related to the Facebook profiles of students in the same class. The students were asked to organize themselves into groups to perform a simulated task. The study results clearly showed that beyond proficiency, strength of friendship and popularity were determinant for the group selection. As further confirmation of this phenomenon, the most skilled users were not always preferred. From another perspective,
Cao et al. [6] exploited URL-sharing information to design a classifier that was able to distinguish between organized and organic groups on Twitter. Their experimental study showed that organic groups’ shared topics were more focused and sharing behaviors were more similar compared with those of organized groups. Another line of research examined group selection for task allocation on crowdsourcing platforms [23, 34, 1]. Li et al. [23] proposed an effective predictive model for identifying accurate working groups based on workers’ attributes relevant to the task characteristics. Rahman et al. [34] noted that beyond worker skills and wages, worker-worker affinity is a key element to consider optimizing collaboration effectiveness. Abraham et al [1] focused on determining the optimal size of a group of workers allowing to achieve a good balance between the cost and quality outcome of a human intelligence task such as label quality assessment. The authors proposed an adaptive stopping rule which decides during the process of worker hiring whether the optimal group size of crowd-workers has been reached. The rule mainly relies on the level of task difficulty and workers’ skills to achieve the task.

In summary, previous work unveils interesting findings among which group performance is affected by three main factors captured at both individual and group levels: (1) members’ knowledge and skills with regard to the task requirements, (2) members’ authority within the social network measuring the users’ expertise and trust towards the task, and (3) group size allowing a good quality-cost trade-off to achieve the collaborative task.

3. PROBLEM FORMULATION AND METHOD OVERVIEW

Based on the literature review highlighting the challenge of solving an IR task in the context of SNQ&A [29] and the potential of collaboration within social networks [17, 28] as well as IR tasks [37, 13], we address the problem of recommending social collaborators for a tweeted question. To do so, we design a task that identifies a group of complementary answerers who could provide the questioner with a cohesive and relevant answer. Taking into account the long-term perspective of gathering users who are willing to collaborate, we carefully consider the problem of group size optimization in terms of balance between the benefits of collaboration and the cognitive cost underlying crowdsourcing [20, 1]. Therefore, for a tweeted question q, the main issue consists in identifying the smallest collaborative group g ⊆ U in which each user u_j ∈ g can contribute to providing a relevant and cohesive answer to the given question q. Accordingly, we believe that the intuition underlying the level at which user u_j contributes given collaborative group g could be assimilated to an information gain-based metric noted as IG(g, u_j). Instead of maximizing the average of information gain over users (which could be seen as a NP-hard problem as mentioned in [34]), we propose the CRAQ model in which, given a posted question q and an initial group U of users, we build the collaborative group g ⊆ U iteratively by discarding users providing the lowest information gain, leading to the maximization of the group entropy IG(g). We provide a general overview of our two-phase CRAQ general methodology in Figure 1. In phase A (Section 4.2), we build a predictive model offline that can predict the likelihood of users’ pairwise collaboration. To attain this goal, we rely on the finding that group performance is affected by two main factors [23, 34]: (1) members’ knowledge and skills relevant to the task requirements, and (2) members’ authority within the social network. With this in mind, we develop a predictive model of a collaborative group based on group members complementarity and authority. In phase B (Section 4.3), given a tweeted question, we first build a group of candidate users who posted topically relevant tweets. Then, considering the predictive model of collaboration built in phase A, we iteratively discard non-informative users and build a smaller group of answerers to recommend to the questioner by ensuring that each user in the group contributes to the collaborative response. This contribution is estimated using the group-based information gain metric. Finally, top K tweets posted from each of the candidate answerers are returned as a whole answer to the user’s question.

4. METHOD

4.1 Preliminaries

Before detailing our method for recommending a group of social collaborators, we introduce the underlying key concept related to the group-based information gain, noted IG(g, u_k). The latter estimates each user’s u_k ∈ g contribution to group g with respect to question q and their complementarity. The metric is estimated as the difference between the group entropy H(g) and conditional entropy H(g|u_k):

\[ IG(g, u_k) = H(g) - H(g|u_k) \]  

These two entropy-based metrics denote the two assumptions related to the topical relevance of users’ tweets and users’ skills within the collaborative group:

- Group entropy H(g) measures the amount of information provided by all users u_j belonging to group g given query q (Eq. 2). Because group entropy is mostly related to the topic of the question, we estimate p(u_j) according to a user-query similarity-based probability p(u_j|q) (Eq. 3):

\[ H(g) = -\sum_{u_j \in g} P(u_j) \cdot \log(P(u_j)) \]  

\[ H(g) \propto -\sum_{u_j \in g} P(u_j|q) \cdot \log(P(u_j|q)) \]

where \( P(u_j|q) \) expresses the probability that user u_j answers question q, as estimated by the normalized cosine similarity between question q and the multinomial representation of the whole set of tweets posted by user u_j.
• Conditional entropy $H(g|u_k)$ expresses the information of a single user $u_k$ in group $g$:

$$H(g|u_k) = p(u_k) \cdot \left( - \sum_{u_j \in g} p(u_j|u_k) \cdot \log(p(u_j|u_k)) \right)$$  \hspace{1cm} (4)

To achieve our objective of building a collaborative group ensuring the authority of each user $u_j$ and his complementarity with respect to each group member $u_{j'}$, we define the collaboration likelihood indicator $L_{jj'}$ estimated between users $u_j$ and $u_{j'}$ (this notion is detailed in Section 4.2). Therefore, we propose to estimate $P(u_j|u_{j'})$ as follows:

$$P(u_j|u_{j'}) = \frac{L_{jj'}}{\sum_{u_k \in g} L_{kk'}}$$  \hspace{1cm} (5)

### 4.2 Learning the Pairwise Collaboration Likelihood

The collaboration likelihood $L_{jj'}$ estimates the potential of collaboration between a pair of users $u_j$ and $u_{j'}$ based on authority and complementarity criteria. We build a predictive model that learns the pairwise collaboration likelihood offline, according to the following statements:

- S1: On Twitter, collaboration between users is noted by the “@” symbol [10, 16].
- S2: Trust and authority enable to improve the effectiveness of the collaboration [26].
- S3: Collaboration is a structured search process in which users might or might not be complementary [36, 37].

More formally, in phase A (Figure 1), we aim to develop a predictive model of the collaboration likelihood using the logistic regression algorithm. In order to train the model, we rely on a set of collaboration pairs $P_{jj'}$ of users $u_j$ and $u_{j'}$ which are modeled according to two elements:

1) The collaboration likelihood $L_{jj'}$ expressed by a boolean indicator characterizing a collaboration initiated by user $u_j$ and directed to user $u_{j'}$. The likelihood value depends on whether the mentioned user $u_{j'}$ provides feedback ($L_{jj'} = 1$) or not ($L_{jj'} = 0$) through an interaction directed at initial user $u_j$. Indeed, according to statement S1, we hypothesize that the likelihood of a pairwise collaboration can be deduced from the social interactions between the users. Moreover, we assume that the collaboration likelihood is valuable if the mentioned user provides feedback (a reply, retweet and/or mention) to the user who initiated the collaboration.

2) The pairwise collaboration $P_{jj'}$ of users $u_j$ and $u_{j'}$ is represented by a set of features $\mathcal{F} = \{f_1, \ldots, f_m\}$ denoting social and collaboration abilities in accordance with statements S2 and S3. Therefore, we consider two categories of features estimated at the pair level:

- **Authority-based features** aim to measure the trust and the expertise of each user (Statement S2).

- **Complementarity-based features** aim to measure the extent to which collaborators are complementary in regards to the SNQ&A task (statement S3). We consider three complementarity dimensions: (1) topicality, which addresses different content-based aspects of the question, (2) the types of information provided (video, links, images, etc.), which offers a wide range of pieces of information, and (3) opinion polarity, which provides contrastive subjective information.

It is worth mentioning that the offline trained predictive model is re-injected within the collaborative group building algorithm (Section 4.3) through the group-based information gain metric detailed in Equation 1.

### 4.3 Building the collaborative group of users

Based on a previous finding highlighting that maximizing the group entropy is equivalent to minimizing the information gain [33], we propose an algorithm (Algorithm 1) in phase B (Figure 1) for recommending a collaborative group. Given an initial group of users, the algorithm discards the least informative user step-by-step with the objective of maximizing the group entropy.

#### Algorithm 1 Social collaborator recommendation

1: **Input:** $C$ : $q$
2: **Output:** $g$
3: $T = RelevantTweets(C, q)$
4: $U = Authors(T)$
5: $t = 0$
6: $g^t = U$

\[ \triangleright \text{Learning the collaborative group of answerers} \]

7: $t^* = \max_{t \in [0, \ldots, |U|-1]} \frac{\partial IG(g^t, u)}{\partial u}$
8: Given $u^* = \arg\min_{u \in g^t} IG(g^t, u)$
9: And $g^{t+1} = g^t \setminus u^*$
10: **return** $g^{t*}$

Where:
- $C$: collection of tweets
- $RelevantTweets(C, q)$: Relevant tweets from collection $C$ considering query $q$
- $Authors(T)$: authors’ of tweets belonging to set $T$

First, we rely on the compatibility of the question topic and the user’s skills, inferred from the user’s tweet topics, to identify the initial group of candidate users. Of note, group initialization is only guided by topical relevance to the question and completely neglects the answerers’ social relationships with the questioner. Because capturing Twitter users’ skills relevant to a topic is a difficult task [31], we rely on the assumption that user relevance may be inferred from the relevance of the his/her tweets. For this purpose, we build the set $T$ of assumed relevant tweets $t_i$ using a tweet retrieval model (e.g., Berberich et al. [4]). The authors of the identified tweets in $T$ are used to build the initial set $U$ of candidate users $u_j$.

Next, we propose an iterative algorithm that decrements the set of candidate users by discarding user $u_k$, who is the least informative user of the group according to the group-based information gain metric (Equation 1). Therefore, for iteration $t$, collaborative group $g^t$ is obtained as follows:

$$g^t = g^{t-1} \setminus u^{t-1}$$  \hspace{1cm} (6)

with $u^{t-1} = \arg\min_{u_{j'} \in g^{t-1}} IG(g^{t-1}, u_{j'})$

where $g^{t-1}$ and $g^t$ represent group $g$ at iterations $t-1$ and $t$, respectively. $u^{t-1}$ expresses the user who has the highest likelihood of being decremented to group $g^{t-1}$, taking into account the group-based information gain criteria.
In identifying the algorithm’s convergence point, we should be able to detect the iteration \( t - 1 \), in which the contribution of user \( u^{t-1} \) to group \( g^{t-1} \) (expressed by \( IG(g^{t-1}, u^{t-1}) \)) is so high that the group would suffer if it did not include the user. One way to confirm this notion is to analyze the difference in information gain when user \( u^{t-1} \) is removed from group \( g^{t-1} \). When this difference is higher than the one obtained for other users who were removed at different timestamps, user \( u^{t-1} \) should be deeply engaged in the collaborative process, and the collaborative group \( g^{t-1} \) should no longer be decremented. At this level, we believe that the collaboration benefits (workers’ skill leveraging) do not outweigh the collaboration costs (cognitive effort, which is positively correlated with the group size, as suggested in [11]). In terms of function analysis, this condition can be assimilated into the iteration \( t \), which maximizes the second partial derivative of the information gain provided by the user who is expected to be removed. Assuming that the information gain is dependent on the group entropy which evolves at each iteration and that the information gain might be biased [33], we used the information gain ratio \( IG_r(g^t, u) \) at each iteration \( t \) which normalizes the metric \( IG(g^t, u) \) by the group entropy \( H(g^t) \), namely
\[
IG_r(g^t, u) = IG(g^t, u)/H(g^t).
\]

The optimization problem might be formulated as follows:
\[
t^* = \arg \max_{t \in [0,...,[T]\{]} \frac{\partial^2 IG(g^t, u)}{\partial u} |_{u = u^t}
\]
\[
\text{Given } u^t = \arg \min_{u \in g^t} IG_r(g^t, u)
\]
\[
\text{And } g^{t+1} = g^t \setminus u^t
\]

where the first partial derivative is estimated by the difference between \( IG_r(g^t, u) \) and \( IG_r(g^{t-1}, u) \) with \( u \) satisfying Equation 9 respectively for group \( g^t \) and \( g^{t-1} \). The second partial derivative is estimated by the differences between the two associated partial derivatives at the first level.

5. EXPERIMENTAL DESIGN

The objective of our evaluation is to measure the impact of our group recommendation model on the effectiveness of a social network question-answering task. For this aim, we consider a SNQ&A setting in which we attempt to evaluate how much the questioner leverages from the collaborative-based answering.

Accordingly, we estimate both the cohesiveness and relevance of the tweets posted by the members of the recommended group. The research questions that guided our evaluation are the following ones:

**RQ1:** Do the tweets posted by the collaborative group members recommended by the CRAQ allow the building of an answer? (Section 6.1)

**RQ2:** Are the recommended group-based answers relevant? (Section 6.2)

**RQ3:** What is the synergic effect of the CRAQ-based collaborative answering methodology? (Section 6.3)

Our evaluation process relies on the CrowdFlower\(^2\) crowd-sourcing platform. Below, we describe our experimental methodology partially inspired by [21].

\(^2\)http://www.crowdflower.com/

5.1 Dataset acquisition and processing

We use two publicly available\(^3\) [38] crisis-related Twitter datasets in our experiments: (1) the Sandy hurricane dataset, referring to the most destructive hurricane in the United States and representing a natural disaster crisis. The Twitter stream was monitored from 29\(^{th}\) to 31\(^{st}\) October 2012 using the keywords sandy, hurricane and storm, and the stream provided a dataset of 4,853,345 English tweets; (2) the Ebola dataset, referring to the virus epidemic in West Africa and representing a public health crisis. This dataset was collected from July 29\(^{th}\) to August 28\(^{th}\) 2014 using the keyword ebola and includes 4,815,142 English tweets. The dataset was cleaned using an automatic methodology proposed by Imran et al. [18] and exploited in [38]. The objective was to distinguish informative tweets that were connected to the crisis or assisted in the understanding of the context from non-informative tweets (those not related to the crisis). We established a three-step process: (1) building a training dataset, including 1,800 tweets manually labeled by 10 human judges; (2) refining the classification model using a logistic regression that relies on 12 features. The latter are divided into three classes and are based on the tweet content features (e.g., the number of hashtags), typography features (e.g., the number of punctuation characters) and vocabulary features (e.g., the number of terms belonging to the dictionary extracted from tweets classified as informative in the training dataset); and (3) dataset filtering, in which we only retained tweets identified as informative. Table 1 presents the descriptive statistics of the resulting datasets.

<table>
<thead>
<tr>
<th>Collection</th>
<th>Sandy</th>
<th>Ebola</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweets</td>
<td>2,119,854</td>
<td>2,872,890</td>
</tr>
<tr>
<td>Microbloggers</td>
<td>1,258,473</td>
<td>750,829</td>
</tr>
<tr>
<td>Retweets</td>
<td>963,631</td>
<td>1,157,826</td>
</tr>
<tr>
<td>Mentions</td>
<td>1,473,498</td>
<td>1,826,059</td>
</tr>
<tr>
<td>Reply</td>
<td>63,596</td>
<td>69,773</td>
</tr>
<tr>
<td>URLs</td>
<td>596,393</td>
<td>1,309,919</td>
</tr>
<tr>
<td>Pictures</td>
<td>107,263</td>
<td>310,581</td>
</tr>
</tbody>
</table>

5.2 Question identification

The first step consists in identifying questions raised on Twitter, with particular attention to identifying questions that are characterized by the intent of obtaining an answer. We apply the question identification methodology proposed in [21] by (1) filtering tweets ending with a question mark; (2) excluding mention tweets and tweets including URLs; (3) filtering tweets with question-oriented hashtags as defined in [20] (e.g., #help, #askquestion, ...), as these hashtags indicate that the questioner is likely to be receptive to answers provided by non-network members [21]; and (4) excluding rhetorical questions. In the last step, we ask CrowdFlower workers to manually annotate questions to determine whether they are rhetorical. Each annotation is redundantly performed by three crowd workers, and a final label is retained using CrowdFlower’s voting formula. For quality control, we include for each task predefined pairs of question and answer as the gold standard for each task. The annotation task cost is 10 cents per question. Of the 87 and 33 question tweets extracted from the Sandy and Ebola datasets, the identification of rhetorical questions leads to 41 and 22 question tweets, respectively. We outline that only one question

\(^3\)https://figshare.com/collections/expac/3283118
obtained a “reply”, which is consistent with previous findings about the lack of answers to tweeted questions [21].

5.3 Evaluation protocol
For evaluation purpose, we adopt a cross-validation methodology in which (a) we learn the predictive model of the pairwise collaboration likelihood (Section 4.2) using one of the datasets and (b) we build the collaborative group (Section 4.3) using the other dataset by estimating the collaboration likelihood metric according to the previously learnt predictive model. The features used are presented in Table 2 and follow the assumptions presented in Section 4.2. Two categories of features are used: the first category aims to measure the authority and the trust of users while the second category measures the complementarity between two collaborators. These two categories are distinguished by the computation point of view. Accordingly, the features in each category are estimated as follows:

- **Authority-based features**: we build a unique metric \(X_{j,j'}\) for a pair of users \(u_j\) and \(u_{j'}\) based on the users’ intrinsic value \((X_j, X_{j'})\), respectively, for users \(u_j\) and \(u_{j'}\) of a particular feature, which is based on their importance (e.g., the number of followers), engagement (e.g., the number of tweets), and activity within the question topic (e.g., the number of topically-related tweets). Thus, we propose to combine the users’ intrinsic value to measure whether these values are in the same range. In other words, we average the metric values of the two collaborators and divide by their standard deviation to reduce the impact of a wide value range. Moreover, to limit the importance of over-socially active users (such as “News” accounts), we transform the mean-standard deviation ratio using the log function:

\[
X_{j,j'} = \log\left(\frac{\mu(X_j, X_{j'})}{\sigma(X_j, X_{j'})}\right)
\]

(10)

- **Complementarity-based features**: we propose to combine each intrinsic value \(X_j\) and \(X_{j'}\), respectively, of users \(u_j\) and \(u_{j'}\) (e.g., the number of tweets with video or tweets with positive opinion) by estimating the absolute difference normalized by the sum of the two collaborators:

\[
X_{j,j'} = \frac{|X_j - X_{j'}|}{X_j + X_{j'}}
\]

(11)

We note that the topical complementarity feature between users is estimated differently. We computed the Jansen-Shannon distance between the topical-based representation of users’ interests obtained through the LDA algorithm.

5.4 Baselines and comparisons
We evaluate the effectiveness of the CRAQ in comparison to five baselines with respect to the following objectives:

- In order to evaluate the effectiveness of our user-driven and collaborative-oriented approach, we compare our proposed model with the MMR, the diversity-based method which ranks tweets by combining relevance and diversity criteria [7]. For fair comparison, we build for each question the top-\(N\) tweets retrieved to the question by setting \(N\) equal to the number of tweets extracted from the collaborative group members recommended by the CRAQ.

- To evaluate the effect of the collaborative group building in the CRAQ, we consider CRAQ-TR, the method that only considers the initial group of candidate collaborators so skipping the group entropy maximization process. To ensure a fair comparison with the CRAQ, we adopt the same approach detailed in Section 4.3, consisting of inferring users from a set of tweets that are ranked using the topic and temporal-oriented tweet ranking model proposed in [4]. We then build collaborative groups with top-ranked users by ensuring that the size of the group resulting from each question equals the size obtained with the CRAQ.

- To evaluate the effect of our proposed collaborative group-based answering approach with respect to an individual user-oriented one, we compare our proposed model with \(U\), the method that only considers the top ranked user provided by the CRAQ-TR [4].

- To assess the contribution of the CRAQ algorithm with regard to state-of-the-art close approaches, we also run the following baselines that include recent related work:

  - **SM**, referring to a structure-based model proposed by [6], in which we use the user group extraction methodology based on URL-sharing behaviors. After the community detection step, we employ the highest cohesion metric value [5] to identify the collaborative group to be recommended. This baseline would enable to evaluate the impact of the social network structure in a group building algorithm.

  - **STM**, referring to the Topic Sensitive Page Rank (TSPR) model that relies on both structure and topic to identify important entities with respect to a query [14]. For our experiments, we represent a user through a multinomial distribution of terms of his/her published tweets and build a user network using mention relationships. The TSPR model is then applied to extract a ranking of users. Similar to CRAQ-TR, for a given question, the final group includes the same number of users as that in the group recommended by the CRAQ. This baseline would enable to evaluate the impact of collaborative assumptions supporting the CRAQ since this baseline only includes structure and topical assumptions in the group building algorithm.

5.5 Evaluation workflow and Ground truth
Two main tasks are performed by the crowd workers and each task is compensated 35 cents per question. For greater reliability, each task is performed by three crowd workers, and the final label is chosen using the CrowdFlower’s voting protocol.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Importance</td>
<td>Number of followers</td>
</tr>
<tr>
<td>Authority</td>
<td>Number of favorites</td>
</tr>
<tr>
<td>Engagement</td>
<td>Number of tweets</td>
</tr>
<tr>
<td>Activity within the topic</td>
<td>Number of topically-related tweets</td>
</tr>
<tr>
<td>Topic</td>
<td>Jansen-Shannon distance between topical-based representation of users’ interests obtained through the LDA algorithm</td>
</tr>
<tr>
<td>Multimedia</td>
<td>Number of tweets with video</td>
</tr>
<tr>
<td>Opinion polarity</td>
<td>Number of tweets with neutral opinion</td>
</tr>
<tr>
<td>Opinion</td>
<td>Number of tweets with only text</td>
</tr>
<tr>
<td>Engagement within</td>
<td>Number of tweets with positive opinion</td>
</tr>
<tr>
<td>Authority</td>
<td>Number of tweets with negative opinion</td>
</tr>
</tbody>
</table>

Table 2: Authority and complementarity-based features modeling collaborators
6. RESULTS AND DISCUSSION

We present here qualitative and quantitative analysis of the CRAQ. We outline that we fixed $N = 100$ (phase A, Figure 1) resulting in groups $U$ with size $|U| \in [25, 94]$, since several retrieved tweets might be posted by the same user. Algorithm 1 provided recommended groups $g \subseteq U$ with size $|g| \in [2, 11]$ with only 1 group including more than 5 collaborators.

### 6.1 Answer building

We start by addressing RQ1 and test whether (1) the CRAQ is effective in providing useful tweets in terms of relatedness to the question topic and complementarity (Table 4) and (2) those tweets allow building a cohesive answer (Table 5).

In Table 4, we see that tweets of users belonging to the groups recommended by the CRAQ are generally related to the topic of the question and are helpful in answering the question. Indeed, the CRAQ obtains the lowest rate for the “Not related” category, with 12.20% and 7.58% for the Sandy and Ebola datasets, respectively, in comparison with the baseline models, where the proportion of “Not related” tweets varies between 12.20% and 50.41% for the Sandy dataset and 9.09% and 66.70% for the Ebola dataset. Moreover, if we aggregate the results obtained for categories 2 and 3 (“2+3 Related and helpful”), the CRAQ also obtains the highest proportion, with 44.72% and 37.88% for Sandy and Ebola, respectively, with values ranging between 26.02% and 44.15% for Sandy and between 12.12% and 34.85% for Ebola ($p < 0.05$ for all baselines). However, in terms of complementarity, we can see that the CRAQ does not obtain higher results than those obtained by the baselines, with rates equal to 20.33% and 10.61% for both datasets, respectively. This result might be explained by the fact that the collaborative feature-based model was learnt (phase A, Section 4.2) from social interactions between pairwise users who were likely in a local neighborhood context and consequently, similar to each other [27, 21]. To gain a deeper understanding of this observation, we computed the significant complementarity features and related regression estimate values using the logistic regression modeling the collaboration likelihood. As can be seen in Table 6, most of complementarity-based features are significant and all the regression estimate values are negative. The latter observation confirms our previous expectation about the low-level of complementarity used as evidence for learning the collaboration likelihood.

However, by analyzing whether crowd workers are able to formulate an answer based on the users’ tweets, Table 5 highlights that this lack of tweet complementarity does not impact on the ability of the recommended group to answer the query. Indeed, the CRAQ achieves one of the highest rates of selected tweets (36.94% for Sandy and 27.42% for
Table 4: Relatedness and complementarity of the CRAQ results. 0: Not related, 1: Related but not helpful, 2: Related, helpful but redundant, 3: Related, helpful, and complementary, 2+3: Related and helpful tweets

<table>
<thead>
<tr>
<th></th>
<th>Sandy dataset</th>
<th>Ebola dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>50.41%</td>
<td>43.90%</td>
</tr>
<tr>
<td>1</td>
<td>23.58%</td>
<td>26.82%</td>
</tr>
<tr>
<td>2</td>
<td>13.82%</td>
<td>17.07%</td>
</tr>
<tr>
<td>3</td>
<td>12.20%</td>
<td>12.06%</td>
</tr>
<tr>
<td>2+3</td>
<td>26.02%</td>
<td>29.16%</td>
</tr>
</tbody>
</table>

Table 5: Answer generation feasibility using the recommended users’ tweets.

<table>
<thead>
<tr>
<th></th>
<th>Sandy dataset</th>
<th>Ebola dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg percentage of selected tweets</td>
<td>14.12%</td>
<td>37%</td>
</tr>
<tr>
<td># built answers</td>
<td>43</td>
<td>29</td>
</tr>
</tbody>
</table>

Table 6: Analysis of the predictive features of collaboration likelihood. −: \( p > 0.05 \), * ***: \( p \leq 0.001 \)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Sandy Regres. estimate</th>
<th>Ebola Regres. estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topical</td>
<td>-</td>
<td>-0.23***</td>
</tr>
<tr>
<td>Images</td>
<td>-</td>
<td>-0.19***</td>
</tr>
<tr>
<td>Links</td>
<td>-0.08****</td>
<td>1.58***</td>
</tr>
<tr>
<td>Hashtags</td>
<td>-0.14***</td>
<td>-0.90***</td>
</tr>
<tr>
<td>Only text</td>
<td>-0.08****</td>
<td>-1.33**</td>
</tr>
<tr>
<td>Positive opinion</td>
<td>-0.12****</td>
<td>-0.16**</td>
</tr>
<tr>
<td>Neutral opinion</td>
<td>-0.16**</td>
<td>-0.16**</td>
</tr>
<tr>
<td>Negative opinion</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Ebola, with \( p < 0.01 \) for MMR, CRAQ-TR, and SM) and the highest number of built answers (77 and 41 for both datasets, respectively). We outline that the U baseline, although characterized by the highest rate of selected tweets (not significantly different from the CRAQ), was not able to select tweets allowing crowd-workers to compose an answer compared to the CRAQ (29 vs. 77, 11 vs. 41, \( p < 0.001 \) for both Sandy and Ebola datasets respectively). For reminder, each pair of question-users’ tweets (respectively 41 and 22 for Sandy and Ebola datasets) has been analyzed by three crowd-workers, leading respectively to 123 and 66 possible built answers for Sandy and Ebola datasets. To summarize, although the CRAQ is not the most effective scenario for selecting complementary tweets, it provides useful information nuggets allowing to build an overall answer to the tweeted question. The key question about the relevance of those answers is addressed in the following section.

### 6.2 Relevance of built answers

Addressing RQ2, Table 7 shows the results of the assessment of the relevance of crowd-built answers based on the tweets provided by the automatically recommended group of users using the CRAQ algorithm and the five baselines. We highlight that \( \chi^2 \)-statistical tests performed on the number of built answer distribution revealed significant differences between the CRAQ and most of the baselines with \( p < 0.001 \); except for SM baseline for the Ebola dataset where \( p > 0.05 \).

We can see from Table 7 that compared to the five baselines, the CRAQ enables to build a higher number of answers, among them a higher proportion of “Partly relevant” and “Relevant” answers. For instance, 77.92% (resp. 75.61%) of the CRAQ answers are considered as relevant while the U baseline obtained 68.97% (resp. 72.72%) for the Sandy dataset (resp. Ebola). This statement reinforces our intuition that a single user might have an insufficient knowledge (even if related) to solve a tweeted question. We outline that although the relevance distribution between the MMR and the CRAQ are relatively similar, the number of built answers remains very low for the MMR (half-lower than the CRAQ). These results give rise to the benefit of building answers from the user’s perspective rather than the tweets regardless of their context.

More particularly, in order to highlight the effect of the CRAQ algorithm, we focus here on the CRAQ-TR which is, moreover, the best baseline in terms of the number of built answers and presents an equivalent number of answers as obtained by the CRAQ algorithm. A correlation analysis between the number of built answers and the relevance of those answers reveals that simply recommending authors of top-ranked tweets (the CRAQ-TR) enables to select tweets related to the question, but mostly not relevant. Indeed, for the Sandy dataset (resp. Ebola), the correlation analysis indicates a non-significant coefficient value of 0.31 (resp. 0.20) for the CRAQ-TR and, unlikely, a significant coefficient of 0.89 with \( p < 0.001 \) (resp. 0.84, \( p < 0.001 \)) for the CRAQ algorithm. In the same mind, we notice that the CRAQ is characterized by a higher number of built answers which have been assessed as “(Partly) relevant” (2+3) than the CRAQ-TR (resp. 60 vs. 52 for Sandy and 31 vs. 24 for Ebola). Intuitively, this difference could be explained by the crowd workers’ assessments of the complementarity and relatedness of group tweets, as the CRAQ obtained a lower rate of non-related recommended groups than the CRAQ-TR (e.g., 12.20% vs. 20.33% for the Sandy dataset). These observations suggest that building a group by gathering individual users identified as relevant through their skills (tweet topical similarity with the question) is not always appropriate. This finding emphasizes the benefit of recommending groups in which users taken as a whole contribute to the group entropy based on their social activity in addition to their skills.

### 6.3 Synergic effect of CRAQ-based collaborative answering methodology

Turning to RQ3, we aim to measure the synergic effect of simulated collaboration within the recommended groups of users with respect to the search effectiveness based on the tweets published by those group members (the CRAQ algorithm and the five baselines). Effectiveness is measured using traditional precision, recall and F-measure metrics. Statistical significance of observed differences between the performance of compared runs is tested using a two-tailed
Table 7: Analysis of the relevance of the crowd-built answers. ba: Number of built answers, 1: Not relevant, 2: Partly relevant, 3: Relevant, 2+3: (Partly) Relevant

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sandy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ba</td>
<td>43</td>
<td>29</td>
<td>75</td>
<td>74</td>
<td>67</td>
<td>77</td>
</tr>
<tr>
<td>1</td>
<td>11: 25.58%</td>
<td>9: 31.03%</td>
<td>23: 30.67%</td>
<td>24: 32.43%</td>
<td>21: 31.34%</td>
<td>17: 22.08%</td>
</tr>
<tr>
<td>2</td>
<td>20: 46.51%</td>
<td>14: 48.28%</td>
<td>33: 44.00%</td>
<td>34: 45.95%</td>
<td>24: 35.82%</td>
<td>39: 50.65%</td>
</tr>
<tr>
<td>3</td>
<td>12: 27.91%</td>
<td>6: 20.69%</td>
<td>19: 25.33%</td>
<td>16: 21.62%</td>
<td>22: 32.84%</td>
<td>21: 27.27%</td>
</tr>
<tr>
<td>2+3</td>
<td>32: 74.42%</td>
<td>26: 68.97%</td>
<td>52: 69.33%</td>
<td>58: 67.57%</td>
<td>46: 68.66%</td>
<td>60: 77.92%</td>
</tr>
<tr>
<td>Ebola</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ba</td>
<td>22</td>
<td>11</td>
<td>39</td>
<td>30</td>
<td>37</td>
<td>41</td>
</tr>
<tr>
<td>1</td>
<td>4: 21.05%</td>
<td>3: 27.27%</td>
<td>15: 38.36%</td>
<td>8: 26.67%</td>
<td>15: 40.54%</td>
<td>10: 24.39%</td>
</tr>
<tr>
<td>2</td>
<td>6: 31.58%</td>
<td>4: 36.36%</td>
<td>18: 46.15%</td>
<td>15: 50%</td>
<td>16: 43.24%</td>
<td>22: 53.66%</td>
</tr>
<tr>
<td>3</td>
<td>9: 47.37%</td>
<td>4: 36.36%</td>
<td>6: 15.38%</td>
<td>7: 23.33%</td>
<td>6: 16.22%</td>
<td>9: 21.95%</td>
</tr>
<tr>
<td>2+3</td>
<td>15: 78.95%</td>
<td>8: 72.72%</td>
<td>24: 1.53%</td>
<td>22: 59.46%</td>
<td>31: 75.61%</td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Effectiveness analysis at the tweet level for recommended collaborative groups. %Chg: CRAQ improvement. Significance t-test: *: 0.01 < α ≤ 0.05, **: 0.001 < α ≤ 0.01, ***: α ≤ 0.001

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sandy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value</td>
<td>%Chg</td>
<td>Value</td>
<td>%Chg</td>
<td>Value</td>
<td>%Chg</td>
<td>Value</td>
</tr>
<tr>
<td>Precision</td>
<td>0.24</td>
<td>+92.93**</td>
<td>0.46</td>
<td>+2.32</td>
<td>0.33</td>
<td>+42.01*</td>
</tr>
<tr>
<td>Recall</td>
<td>0.09</td>
<td>+95.19*</td>
<td>0.1</td>
<td>+81.63*</td>
<td>0.15</td>
<td>+16.18</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.12</td>
<td>+78.22*</td>
<td>0.15</td>
<td>+41.79</td>
<td>0.19</td>
<td>+10.59</td>
</tr>
<tr>
<td>Ebola</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value</td>
<td>%Chg</td>
<td>Value</td>
<td>%Chg</td>
<td>Value</td>
<td>%Chg</td>
<td>Value</td>
</tr>
<tr>
<td>Precision</td>
<td>0.22</td>
<td>+153.65**</td>
<td>0.64</td>
<td>-12.12</td>
<td>0.5</td>
<td>+12.22</td>
</tr>
<tr>
<td>Recall</td>
<td>0.07</td>
<td>+155.59**</td>
<td>0.11</td>
<td>+69.96*</td>
<td>0.22</td>
<td>-18.08</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.09</td>
<td>+164.17***</td>
<td>0.21</td>
<td>+17.46</td>
<td>0.28</td>
<td>-11.64</td>
</tr>
</tbody>
</table>

The CRAQ algorithm significantly outperforms the STM for both datasets with respect to the recall and F-measure metrics. This baseline achieves very small recall values (0.1 and 0.06 for the Sandy and Ebola datasets, respectively), suggesting that topically relevant tweets issued from the most socially authoritative are not obviously relevant to answer the tweeted question. This finding reinforces our driving idea around collaborative-based answering and is consistent with previous findings which claim that collaborative groups might be built based on user’s skills regarding the task at hand and user-user affinity likelihood in order to ensure the synergic effect of collaboration [34, 35].
book and community Q&A sites. One limitation is that the predictive model of collaboration likelihood relies on basic assumptions of collaboration (e.g., mention). A deeper analysis of collaboration behavior on social networks would help to identify collaboration patterns, and then better estimate the collaboration likelihood. The relevance ratings of both tweets posted by the CRAQ-based recommended users and the related manually built answers are very promising, and we plan to extend this work through the automatic summarization of candidate answers.

8. ACKNOWLEDGMENTS

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

9. REFERENCES


