Deep Twitter Diving: Exploring Topical Groups in Microblogs at Scale

Parantapa Bhattacharya
IIT Kharagpur, India
MPI-SWS, Germany

Saptarshi Ghosh
BESU Shibpur, India
MPI-SWS, Germany

Juhi Kulshrestha
MPI-SWS, Germany

Mainack Mondal
MPI-SWS, Germany

Muhammad Bilal Zafar
MPI-SWS, Germany

Niloy Ganguly
IIT Kharagpur, India

Krishna P. Gummadi
MPI-SWS, Germany

ABSTRACT

We present a semantic methodology to identify *topical groups* in Twitter on a large number of topics, each consisting of users who are experts on or interested in a specific topic. Early studies investigating the nature of Twitter suggest that it is a social media platform consisting of a relatively small section of elite users, producing information on a few popular topics such as media, politics, and music, and the general population consuming it. We show that this characterization ignores a rich set of highly specialized topics, ranging from geology, neurology, to astrophysics and karate – each being discussed by their own topical groups. We present a detailed characterization of these topical groups based on their network structures and tweeting behaviors. Analyzing these groups on the backdrop of the common identity and bond theory in social sciences shows that these groups exhibit characteristics of topical-identity based groups, rather than social-bond based ones.

Author Keywords

Topical groups; identity-based groups; Twitter; topical experts; seekers of topical information;

ACM Classification Keywords

H.3.5 On-line Information Services: Web-based services; J.4 Computer Applications: Social and behavioral sciences

INTRODUCTION

What is Twitter? Who says what to whom on Twitter? Recently these two questions about the fundamental nature of the Twitter microblogging site have attracted a lot of research attention. Early studies that attempted to answer these questions [16, 30] concluded that Twitter is more like a news media site than a social networking site, and that a small number of elite users consisting of celebrities, bloggers, media representatives, and organizations dominate and control the production and flow of information, which is then consumed by the masses of ordinary users in Twitter. These characterizations of the Twitter network have fostered the current popular perception of Twitter as a platform for getting information or public opinions on topics of general interest.

In this paper, we argue that it is time to revisit these two foundational questions about Twitter. Recently a lot of research has focused on mining user data to develop detailed characterizations of Twitter users. For instance, in our prior works [12, 25], we have proposed techniques to infer the biographical information and topical expertise of millions of Twitter users. These recent advances in accurately profiling large numbers of Twitter users enable us to dive deep into the Twitter network and explore several tens of thousands of *topical groups* that the earlier studies investigating these questions [16, 30] were oblivious to. Below, we summarize the three primary contributions of this paper.

1. Inferring topical groups at scale: A topical group is a collection of users that are interested in propagating or seeking information on a specific topic (e.g., music, politics, environment, neurology, forensics). In this paper, we propose a new *semantic approach* for detecting topical groups in Twitter. Our methodology first identifies *experts*, i.e., authoritative users, on a particular topic and then identifies *seekers*, i.e., followers of the experts on the topic. Unlike prior approaches, our method scales very well: using detailed data about 38.4 million Twitter users, we identify several thousand topical groups covering a wide variety of topics that Twitter users are interested in. The topical groups we identified cover 49.5% of all the users and 94.3% of all the links in the Twitter network for these 38.4 million users.

2. Characterizing the topical groups: We conducted a detailed characterization of the topical groups, focusing on their

CSCW'14, February 15–19, 2014, Baltimore, Maryland, USA.

ACM 978-1-4503-2540-0/14/02 ...$15.00.
http://dx.doi.org/10.1145/2531602.2531636

diversity, the users followed by them, and tweeting behaviors of these users.

(a) Diversity in topical groups: We discover a rich and diverse set of highly specialized and focused topical groups that lie embedded deep within the Twitter network. We find that these topical groups span a variety of niche topics, which range from 'geology' and 'neurology' to 'astrophysics' and 'karate'. These groups are often ignored by research studies today because (i) they are small in size, i.e., the number of group members is low, (ii) even the experts in these niche topical groups have relatively low overall popularity in the global network, and (iii) tweets relevant to the specific topics of interest of the groups are largely limited to the group members. Earlier studies that have attempted to characterize the Twitter network at-scale by looking at popular users or popular tweets have mostly overlooked the existence of these groups. Our work offers a different perspective on what the Twitter network is – a treasure-trove of diverse topical groups.

(b) Following and tweeting behaviors in topical groups: We also analyzed the *following* behavior of experts and seekers within a topical group, and the tweeting behavior of experts within a topical group. Our analysis provides several interesting insights: (i) within a topical group, the experts and seekers exhibit very different connectivities; the experts tend to be considerably more interconnected amongst themselves, while the seekers mostly connect to the experts and not amongst themselves, (ii) experts of niche topical groups, i.e., groups with a few tens to a few hundreds of experts, tend to be particularly tightly interconnected amongst themselves, (iii) the two-hop (and not one-hop) neighborhoods of experts include the majority of experts in the group for both niche and popular groups, (iv) experts, especially those in niche topical groups, tend to tweet on the topics of their expertise, and (v) tweets that are popular amongst experts in a topical group have a high likelihood of being related to the topical expertise of the group.

Overall, the observed behaviors of topical groups match closely with those expected of *identity-based groups* (rather than bond-based groups) in social sciences literature [22, 23].

3. Implications of the findings: Our findings about topical groups have important implications for researchers studying Twitter or designing various services on Twitter. We show that because of the relatively low interconnectivity between experts and seekers, it is difficult, if not impossible, to detect these topical groups by applying standard community-detection techniques on the Twitter network graph. We also demonstrate the significance of our findings by applying them in the context of two different applications: (i) Finding more experts in a topical group: we show how the tight interconnectivity between experts, especially in the case of niche topical groups can be leveraged to discover new (and otherwise missing) group members within the Twitter network, and (ii) Topical content recommendations: we show how the observation that experts in a group tend to tweet on topics relevant to the group can be used to build topical search and recommendation systems.

RELATED WORK

The present study focuses on identifying topical groups in Twitter and understanding their characteristics. There have been several prior studies on detecting and distinguishing between different types of user groups in social networks, in general, and Twitter, in particular. In this section, we discuss some of the major studies in each of these categories.

User groups in social sciences: The topic of user groups (or communities) in societies has been widely studied in behavioral sciences, and a variety of theories have been proposed to explain the formation of such groups [4, 8, 21, 29]. One of the most well-known theories on group formation, both in off-line and online societies, is the *common identity and common bond theory* [22, 23], according to which, the attachment of users to a particular group can be either identity-based or bond-based. Identity-based attachment holds when people join a group based on their interest in the group as a whole or in a well-defined common theme (topic) shared by the members. On the other hand, *bond-based* attachment is driven by personal social relations (bonds) among the members, and may be characterized by the absence of any central theme / topic of discussion within the group. According to this theory, the two types of groups differ in reciprocity of the interactions and the topicality of the discussions. Bond-based groups tend to have higher reciprocity among the members, whereas interactions in identity-based groups are generally not directly reciprocated. Also, the topics of discussion in bond-based groups tend to vary widely and cover multiple subjects, while in identity-based groups, they tend to be related to the common group theme.

There have been several attempts to distinguish between bond-based and identity-based groups, especially in the online world. For instance, [24] classified chats among users on a text-based communication platform into two categories – on-topic chats which centered around a common topic (examples of identity-based groups) and off-topic chats where people chatted on a variety of topics (examples of bond-based groups). More recently, [13] differentiated between the identity-based and bond-based groups (which they called topical groups and social groups respectively) in the Flickr social network.

In this study, we analyze the interconnections, interactions, and tweeting behaviors of *topical groups* in the Twitter social network. Our findings about the behaviors of users in the different topical groups conform to those one would expect to find in identity-based groups.

Detecting groups in social networks: Automatically identifying user groups in large social networks is a topic that has received a lot of recent research attention. A large number of algorithms have been proposed to identify user groups by detecting graph communities in social networks. These community-detection algorithms rely only on the network structure and identify groups of densely connected nodes in the network as communities; see [11] for a detailed review on such algorithms. The communities detected by such algorithms are known to mostly resemble bond-based groups [13]. In fact, the recent study [31] showed that communities iden-

tified from the network structure often do not conform to ground-truth groups that are explicitly defined by the nodes (members) in the network.

Detection of identity-based groups usually requires knowledge of groups defined explicitly by the nodes [13] or some semantic methodology based on the content generated in the OSNs (e.g., [24] analyzed the topicality of the conversations / chats among members). Unfortunately, not all social networks (for example, Twitter) allow users to form explicitly defined groups. In such systems, one would have to resort to semantic approaches for detecting identity-based groups.

In this study, we propose a semantic methodology to detect topical groups of Twitter users who are experts on or interested in a specific topic. We show that the semantically meaningful topical groups resemble identity-based groups, and are very different from the groups detected by standard community detection algorithms (which are likely to be bond-based).

Detecting groups in Twitter: There have been several efforts to identify groups of users having similar interests / characteristics in Twitter. They fall into two broad categories: (i) *graph-based approaches*: those that rely primarily on applying community detection algorithms on the Twitter follow-network [14, 28], and (ii) *semantic approaches*: those that rely on identifying groups of users with similar interests by analyzing either their profile meta-data (e.g., profile names, bios, URLs) or the content of their tweets [2, 3, 9, 30].

The graph-based approaches are scalable, but they would not be able to detect loosely connected groups of users (identity-based groups) driven by some common topical interest. The current semantic methods are better at identifying such groups, but they suffer from certain other drawbacks. First, studies using semantic methods are often limited to a few pre-selected (and popular) topics. Second, many users do not provide information about their topical interests in their profiles, and even if they did, it is unclear if they can be trusted. Similarly, users' tweets often contain conversation on day-to-day activities, making it difficult to identify meaningful topics from tweets [25, 27]. As a result, studies often focus on a few specific and popular topics. For instance, [2] focused on tweets posted by media sources on news, technology, sports, music, politics, business and fashion, while [9] characterized the users who tweet about certain events related to popular topics such as politics (Egyptian revolution, Wikileaks), music and entertainment (Academy Awards, Bonnaroo), environment (Earth Day), sports (Super Bowl) and technology (Release of ipad2).

In this paper, we propose a new semantic approach that scales well to identify several tens of thousands of identity-based groups on a diverse set of topics, covering several tens of millions of Twitter users. The only similar study we know of is [3], which studies the diffusion of popular as well as less popular topics identified using NLP techniques on tweets. However, the methodology used in the present study is very different, which can result in important differences in the topical groups identified. [3] identified topics from tweets and studied the networks among users who tweet on a common topic – so their topical groups include *ephemeral users*, who might have once tweeted about a topic, but do not have a long-standing interest in it. For example, many users might tweet about an election or a environmental disaster the day after the event, even if they have no specific interest in politics or environment. Such ephemeral users would be included in topical groups identified by [3], but not by our methodology.

User roles in spreading information in Twitter: There have been several studies on how information and news related to specific topics spread in Twitter [17, 19], which have classified the different roles played by users in this diffusion. For instance, [30] analyzed the flow of information from 'elite' users to their followers. [9] characterized three types of users – organizations, journalists / media bloggers and ordinary individuals – who post tweets about certain events. In this study, we distinguish between two types of users within a topical group – *experts* who are likely to be authoritative sources of information on specific topics, and *seekers* who are interested in gathering information on these topics.

INFERRING TOPICAL GROUPS

In this section, we discuss our methodology for inferring topical groups in Twitter. By topical groups, we refer to the set of Twitter users who are interested in propagating or receiving information on specific topics. Intuitively, our methodology for identifying topical groups consists of the following steps. First, we identify *experts*, i.e., authoritative users on a specific topic. Second, we find *seekers*, users interested in receiving information on that topic, by looking for users who follow multiple experts on the topic. Third, we group the experts and seekers of information on a specific topic into a topical group.

The primary challenge in our methodology lies in accurately discovering expert Twitter users on a wide variety of topics that might be of interest to Twitter users. For this, we utilize a semantic methodology for inferring topics related to individual users, which scales well with respect to both the variety of topics identified as well as the number of users for whom topics can be inferred. Our methodology relies on the Lists feature in Twitter.

List-based characterization of users

Lists are an organizational feature in Twitter, which users create to group experts on topics that interest them [20]. To create a List, a user specifies a name and an optional description for the List; the user can then add other users as members of the List. Table 1 shows a few examples of Lists, including their names, descriptions and some of their members. The key observation is that the List names and descriptions offer semantic cues to the expertise of the users included in them.

In our prior works [12, 25], we have shown that List names and descriptions can be leveraged to infer the topical characteristics of users. Specifically, we showed that List meta-data can be used to (i) accurately identify the topical characteristics for millions of Twitter users [25], and to (ii) find relevant experts on several tens of thousands of topics [12]. We leverage these prior works, and utilize Lists to identify experts on a wide variety of topics.

List Name	List Description	Members
News	News media accounts	nytimes, BBCNews, WSJ, cnnbrk, CBSNews
Music	Musicians	Eminem, britneyspears, ladygaga, rihanna, BonJovi
Tennis	Tennis players and Tennis news	andyroddick, usopen, Bryanbros, ATPWorldTour
Politics	Politicians & people who talk about them	BarackObama, whitehouse, nprpolitics, billmaher
Geology	Geology, Geophysics, Professionals and Students	Am Geophysical Union, Andy Fyon [Director, Ontario Geological Survey]
The Brain / Neurology	Information related to brain and neuro health	Neuroscience News, SarahJayneBlakemore [Cognitive Neuroscience Professor, UCL]
Forensics	Computer forensics, cyber forensics, phone forensics	Security Tube, Chad Tilbury [network security professional]
Chemistry	Resources on chemistry and its subfields	Clinical Chemistry, Anne Helmenstine [About.com writer on Chemistry]

Table 1: Examples of Lists: their names, descriptions, and some sample members. For less well-known experts, extracts from their Twitter account bio are also given within square braces.

The List-based methodology of identifying topical attributes of users has two key advantages. First, several recent studies [25, 27, 30] have shown that Lists help to infer topical expertise of users much more accurately, as compared to the contents of the tweets posted by the users. Second, since the List-based methodology relies on crowd-sourced social annotations (i.e., on Lists created by Twitter users themselves), it scales very well with respect to both the number of topics identified and the number of users for whom topics of expertise (and interest) can be identified. Note, that unlike most prior studies [2, 9, 30], we did not pre-select a small set of topics or events and identify users related to those topics. Instead, our List-based methodology automatically identifies experts on *any topic on which Twitter users have created Lists, i.e., any topic that interests Twitter users.* As we discuss in the next section, our methodology not only covers a large and diverse set of topics, but it also identifies hundreds of thousands of users who are experts on these topics.

Twitter data gathered

As stated above, we identify topical groups consisting of users who are experts or knowledge seekers on specific topics. Ideally, one would like to gather data about all users presently in Twitter and identify experts and seekers on any topic. However, given the restrictions on using the Twitter API for crawling, it is infeasible for us to gather data about the enormous number of users presently in Twitter (more than 550 million, as of May 2013 [1]). We started crawling users in Twitter network in the sequential order of their account creation date (using the Twitter API). We were able to gather detailed information of about 38.4 million users, including their profile information, their social links (followings), and the Lists in which they are members.

Using the data of the social links and the Lists, we constructed the Twitter *subscription network* in which there is a link from user (node) u to user v if u subscribes to the tweets posted by v. Note that in Twitter, user u can subscribe to v's tweets by either following v or by creating or subscribing to a List containing v as a member.

We also collected the tweets posted by all the users whom we identify as experts during the month of December 2012. Below, we describe how we identified topical experts and seekers in this dataset.

Identifying experts on a topic

To identify topics of expertise of an individual user u, we use the methodology proposed earlier by us in [12, 25] – collect the Lists which have u as a member, and extract the most common terms (topics) that appear in the names and descriptions of the Lists. Specifically, we identify a user u as an expert on some topic T if and only if u has been listed on T at least 10 times, i.e., topic T appears at least 10 times in the names or descriptions of the Lists containing the user. The threshold 10 is selected based on our observations in [12, 25]. Even after applying this threshold, we found that some popular Twitter users can be listed as experts on hundreds of topics. To limit the focus to the primary topics of expertise of an individual user u, we rank all of u's topics of expertise based on the number of times u is listed on each topic, and select the top 50 topics as u's primary topics of expertise. Similar to [12, 25], we considered as topics only unigrams (single words such as 'politics', 'music') and bigrams (two words which frequently occur together, e.g., 'social media', 'video game', 'bay area').

Using this List-based methodology, we identified *584,759 topical experts* on a large diverse set of topics, from among the 38.4 million users whose data we could gather. Table 2 shows some example topics and some of the Twitter users identified as experts on the topic, using this methodology. The table also gives the number of followers (indegree) of the experts in our Twitter subscription network. Note that the identified experts include not only globally popular users having millions of followers (e.g., Barack Obama, Lady Gaga), but also less popular ones having a few hundred to a few thousand followers. (For the less known experts, we also give extracts from their Twitter account bio in Table 2.)

Identifying seekers of topical information

Once we identify experts on a topic, identifying seekers of information on the topic (i.e., users interested in gathering information on the topic) is relatively straight-forward. Intuitively, if a user u subscribes to tweets from several experts on a certain topic, then u is likely to be interested in that topic. We considered u to be interested in topic T if and only if u subscribes to at least K experts on topic T. We experimented with varying values of K= 3, 5, and 10. While the number of seekers falls sharply as K increases, the high-level insights from the analysis using different values of K remain the same. So unless stated otherwise, the default value of K for the results presented in this paper is 3. Similar to the case of identifying topics of expertise, we found that many users in Twitter subscribe to experts on hundreds of topics. To focus on the primary topics of interest of an individual user u, we rank the topics of interest of u based on the number of experts she subscribes to on each topic, and then limit ourselves

Topic	Some experts identified by List-based method
Music	Lady Gaga (1.5M), coldplay (1.3M), Katy Perry (1.2M), Justin Timberlake (904K), Dallas Martin [VP, Warner Bros Records] (750), TenorRyan [Opera singing road warrior] (557)
Politics	Barack Obama (2.1M), Al Gore (1.1M), NPR Politics (976K), John McCain (887K), Bill Maher (285K), BristolRed [Chair of Bristol Labour Party] (631), Scott Fluhr [Harrison County GOP Chairman] (371)
Environment	TreeHugger.com (51K), GreenPeace USA (16K), Dennis Dimick [environment editor @natgeomag] (1328)
Neurology	Oliver Sacks (7K), Neurology Today (1K), MNT Neurology News (591), AAN Public (447), Neurology Journal (296)
Chemistry	ACS Pressroom [News from the American Chemical Society] (1173), Chemistry News (1108), chemheritage [Chemical Heritage Foundation Library] (856), Chemical Science (442)
Forensics	SANS Institute (4K), ForensicFocus (998), Michael Murr [Forensic Scientist] (618), Forensic Archaeology (265), Simon Whitfield [Digital Forensic Investigator] (144)
Geology	geosociety (1107), Kim Hannula [Structural geology professor] (382), Garry Hayes [Teacher of Geology] (232), Dave Mayer [Grad Student, Geoscientist] (158)

Table 2: Examples of topical experts identified by our List-based methodology [12, 25] for specific topics. Examples are shown for popular topics (music, politics, environment) as well as niche topics (neurology, chemistry, forensics, geology). For some of the less known experts, extracts from their Twitter account bio are also given within square braces. The numbers in parentheses give the approximate indegree of the expert in our Twitter subscription network (K: thousand, M:million).

to at most the top 50 topics (i.e., the 50 topics on which u subscribes to most experts).

Note that the set of experts and the set of seekers for a topic are not necessarily disjoint – a user can be both an expert and a seeker on the same topic. In fact, we find that a majority of experts on many topics are also seekers of information on their own topics of expertise.

We use the term 'topical group' to refer to the combined set of experts and seekers on a topic. Once we identify the topical group for a topic T, we can also construct the *topical-network* corresponding to T, by extracting the subgraph of the Twitter subscription network induced by the set of experts and seekers on topic T. In the subsequent sections, we analyze the properties of topical groups and topical networks corresponding to some specific topics.

Scalability of our approach

One of the primary advantages of the above approach is its scalability, with respect to the diversity in the topical groups identified and the number of users whose topical characteristics (expertise / interest) could be identified. The diversity in the topical groups will be studied in detail in the next section. Here, we demonstrate the scalability of the approach in characterizing a very large part of the Twitter network (especially the most popular / important part), by estimating the aggregate impact of all the topical groups that we detected.

We find that as many as 49.5% of all nodes and 94.3% of all links in the Twitter subscription network (for the 38.4 million users) are included in at least one of the topical groups that we extracted. While the experts across the different groups constitute only a small fraction of all users in the Twitter network (less than 1.5%), they include a large majority of the most popular users.

Figure 1 shows the fraction of all users having k or more followers in the Twitter, who are experts in some topical group. It is seen that more than 60% of all users having $k = 1000$ or more followers are included in the set of experts on some topic, and this fraction increases rapidly for higher values of k. Note, however, that *not* all experts have large numbers of followers; in particular, several of the experts in the niche topical groups are only moderately popular. Also, 65.2% of all links in the Twitter network are used to follow the experts.

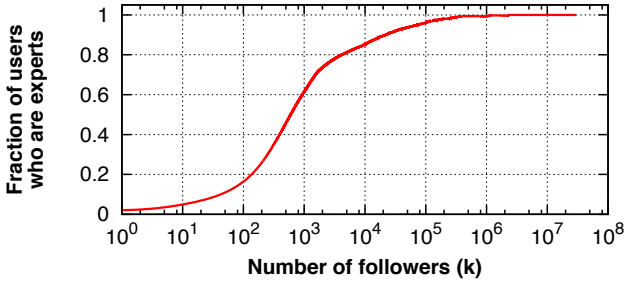


Figure 1: Fraction of all users having k or more followers who are experts – shown for all values of k.

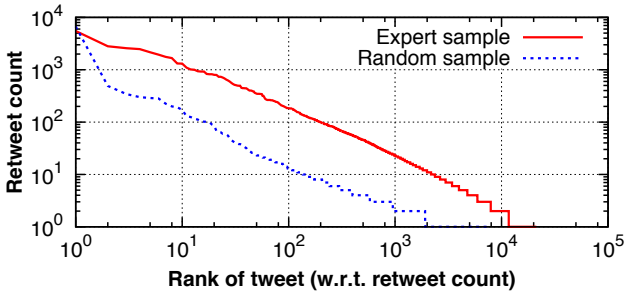
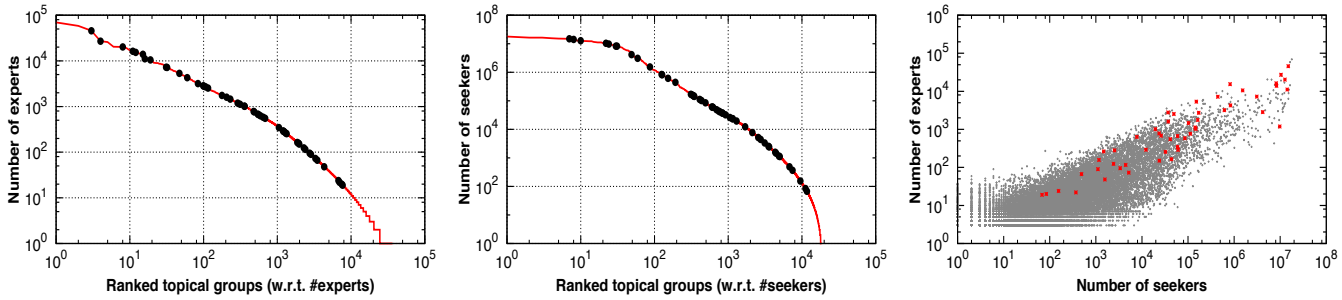


Figure 2: Distribution of the number of times a tweet is retweeted, for tweets posted by experts and tweets in the Twitter random sample. Tweets posted by experts are retweeted many more times than tweets in random 1% sample.

Hence, even though only a small fraction of all Twitter users are experts, a disproportionately large fraction of all follow links in Twitter point to them. Thus their aggregate followers (the seekers) include a substantial fraction of all users in the Twitter network.

Moreover, the content (tweets) posted by the experts constitute the most popular information in Twitter. The popularity of a given tweet in the Twitter network can be estimated by the number of times it has been retweeted in Twitter (also known as the retweet count of the tweet). We collected 100,000 tweets posted by the experts from diverse topical groups, and obtained the retweet counts for these tweets, using the Twitter API. As a baseline for comparison, we also considered 100,000 random tweets, randomly selected from a



(a) Number of experts **(b) Number of seekers** **(c) Scatter plot of experts and seekers**

Figure 3: Distribution of the number of (a) experts and (b) seekers in various topical groups. Figure (c) shows a scatter-plot where the point (x, y) indicates a topical group having x seekers and y experts. Also indicated are the 50 topical groups that are studied in detail in subsequent sections (shown as black circles in (a) and (b), and red asterisks in (c).

1% random sample of all tweets made available by Twitter.[1] The popularity (retweet count) distributions of the random tweets and the experts' tweets is shown in Figure 2. The plot clearly shows that the tweets authored by the experts have an order of magnitude higher retweet counts than random tweets in Twitter. Thus the content posted by the experts is far more popular, and is retweeted numerous times by the seekers who follow the experts.

Taken together, the above discussion shows that the topical groups detected by the List-based approach constitute a significant portion of the Twitter network, thus highlighting the scalability of the approach.

CHARACTERIZING TOPICAL GROUPS

In this section, we present a detailed characterization of the topical groups in the Twitter network. We begin by exploring the diversity in the topical groups, both in terms of the topics they cover as well as their membership (i.e., the number of experts and seekers in the different groups). We then analyze the following and tweeting behaviors of the members of the groups, especially the experts.

Diversity in topical groups

As described earlier, the List-based methodology enables us to identify topical groups potentially on any topic on which a threshold number (10) of Twitter users created Lists. To investigate the diversity in the topics covered by the groups, we only considered those topics that correspond to words appearing in an English dictionary (and some very popular abbreviations of English words, such as 'celebs' and 'tech'). Note that by limiting our analysis to English language topics, we are potentially underestimating the diversity in topical groups.

Figures 3(a) and (b) respectively show the distribution of the number of experts and seekers in various topical groups, where the groups have been ranked in decreasing order of the number of experts and seekers respectively. The plots show that the number of experts and seekers in different topical groups vary by several orders of magnitude. We also checked the correlation between the number of experts and seekers in a group. Figure 3(c) shows a scatter plot of these two measures, where the point (x, y) indicates a topical group having x seekers and y experts. The plot shows that as the number of experts in a topic increases, the number of seekers on that topic also tends to increase. In fact, the number of experts and seekers are highly correlated, with a Pearson correlation coefficient of 0.7 across all topics. Thus, a small number of groups have tens of thousands of experts and hundreds of thousands of seekers interested in the corresponding topics, while there are a large number of groups that each have a few tens or hundreds of experts and a few hundred to a few thousand seekers.

As we shall see in the rest of this section, it helps to distinguish between the topical groups with a large numbers of experts and seekers and those with few experts and seekers. We consider a topic/group as *popular* if it has more than 1000 experts and more than 100,000 seekers. On the other hand, *niche* topics / groups are the ones containing fewer than 500 experts and 5000 seekers. Though these thresholds are somewhat arbitrary and can be altered, the high-level insights drawn in the study remain valid even if the thresholds are varied slightly. Note that our description of popular and niche groups is based on the number of experts and seekers identified from the Twitter data we gathered, and may not always agree with the corresponding notions in the off-line world. For instance, a certain topic may be generally discussed by many people in the off-line world, but if there are only a few hundred to thousand experts and seekers in Twitter on this topic, it would be considered as a niche topic in our study.

To explore the wide diversity of the topics of interest to the topical groups in Twitter, we partitioned the topical groups among several ranges of the number of experts and seekers, and show example topics / groups in each range in Table 3. The table illustrates the rich diversity of topics on which information is shared over the Twitter platform.

The cells in the bottom-right corner of Table 3 show the *popular topics* (in blue) having more than 1,000 experts and more than 100,000 seekers. These are the topics for which Twitter is generally known today, such as 'music', 'technology', 'politics', 'bloggers', 'celebrities', 'fashion', 'hollywood', and so on. As we move towards the top-left cornor of the table, both the number of experts and seekers decrease and topics become more specialized, like 'medicine', 'astronomy', 'yoga',

[1]Twitter provides a 1% random sample of all tweets in real-time, which we collected during December 2012.

No. of seekers	Number of experts					
	< 100	100 – 500	500 – 1K	1K – 5K	5K – 10K	> 10K
< 1K	**(5416)** *geology*, *karate*, *malaria*, *neurology*, *tsunami*, psychiatry, radiology, pediatrics, dermatology, dentistry	**(132)** volleyball, philosophers, tarot, perfume, florists, copywriters, taxi, esperanto				
1K – 5K	**(915)** *biology*, *chemistry*, *swimmers*, astrophysics, multimedia, semiconductor, renewable-energy, breast-cancer, judaism	**(428)** *painters*, *astrology*, *sociology*, *geography*, *forensics*, anthropology, genealogy, archaeology, gluten, diabetes, neuroscience	**(17)** architects, insurance, second-life, police, progressives, creativity			
5K – 10K	**(166)** *malware*, gnu, robot, chicago-sports, gospel-music, space-exploration, wall-street	**(202)** horror, agriculture, atheism, attorneys, furniture, art-galleries, ubuntu	**(34)** *psychology*, poetry, catholic, hospitals, autism, jazz	**(2)** coffee, dealers		
10K – 50K	**(174)** ipod, ipad, virus, Liverpool-FC, choreographers, heavy-metal, backstreet-boys, world-cup,	**(312)** *olympics*, *physics*, *theology*, *earthquake*, opera, makeup, Adobe, wrestlers, typography, american-idol	**(146)** *tennis*, *linux*, *astronomy*, yoga, animation, manga, doctors, realtors, wildlife, rugby, forex, php, java,	**(67)** *law*, *history*, *beer*, *golf*, librarians, theatre, military, poker, conservatives, vegan		
50K– 100K	**(7)** bbc-radio, UK-celebs, christian-leaders, superstars	**(61)** *hackers*, *programmers*, bicycle, GOP, fantasy-football, NCAA, wwe, sci-fi	**(35)** *medicine*, *cyclists*, investors, recipes, NHL, xbox, triathlon, Google	**(37)** hotels, museums, hockey, architecture, charities, weather, space		
> 100K	**(3)** headlines, brits	**(49)** pop-culture, gospel, BBC, reality-tv, bollywood	**(58)** *religion*, actresses, gadgets, graphic-design, directors, lifestyle, gossip, commentators, youtube	**(140)** *books*, *government*, *comedy*, *environment*, *baseball*, *soccer*, *hollywood*, *iphone*, *economics*, money	**(25)** *fashion*, *education*, *wine*, photography, radio, restaurants, science, SEO	**(17)** *music*, *tech*, *business*, *politics*, *food*, *sports*, *celebs*, *health*, media, bloggers, travel, writers

Table 3: Examples of topical groups on different topics, along with the number of seekers and number of experts in each group. The total number of groups having a certain number of experts and seekers is indicated in each cell. The popular topics are the ones shown in blue color (bottom-right corner) and the niche topics are the ones shown in red (top-left corner). The topical groups chosen for detailed study in later sections are *italicized*.

'law', 'history', and 'psychology'; we refer to these as the 'intermediate topics'. The topics in this range also include *technological topics* such as 'hackers', 'linux', 'xbox', 'php', 'java' and 'Google', and *sport-related topics* such as 'golf', 'NHL', 'triathlon', 'hockey' and 'wrestlers'. Finally, the top-left corner of Table 3 contains the *niche topics* (shown in red) having fewer than 500 experts and fewer than 5000 seekers. Most of these topics relate to advanced scientific and technological disciplines like 'biology', 'chemistry', 'dermatology', 'astrophysics', 'geography', 'forensics', 'neuroscience', and 'diabetes'. The table shows that apart from the popular topics, Twitter also contains a collection of groups on a diverse set of highly specialized topics.

In addition to some illustrative topics, each cell in Table 3 also shows the number of topical groups with the number of experts and seekers in the corresponding ranges. We see that the number of niche topics is several orders of magnitude higher than the number of popular topics. For instance, there are more than 6000 niche topics having less than 500 experts and less than 5000 seekers, as compared to about 200 popular topics having more than 1,000 experts and 100,000 seekers. Most of the prior studies on Twitter have focused exclusively on the popular topics [30, 2, 14] and the presence of a large number of niche topical groups has largely been ignored in the literature.

Why niche topical groups are often overlooked

There are several potential reasons why niche topical groups are often overlooked by studies characterizing the Twitter social network. First, note that most niche topical groups have only a small number of experts and users interested in the topics. Second, many of the experts in the niche topical groups are not globally popular. For instance, experts in niche topics like 'neurology' or 'anthropology' rank quite low in global rankings of user influence (computed using metrics such as number of followers or PageRank over the follow network [26]). We computed the global PageRanks (over the Twitter follow network of 38.4M users) of the top experts in different topical groups and compared them. We found while the top experts in popular topical groups are ranked very highly (amongst the top 10 or top 100 users), the experts in niche topical groups are almost always ranked outside the top tens of thousands (or, in some cases, even outside the top hundreds of thousands) of users.

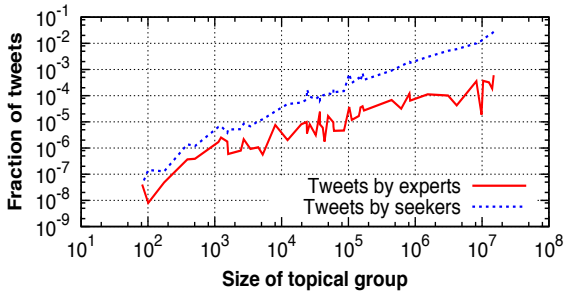
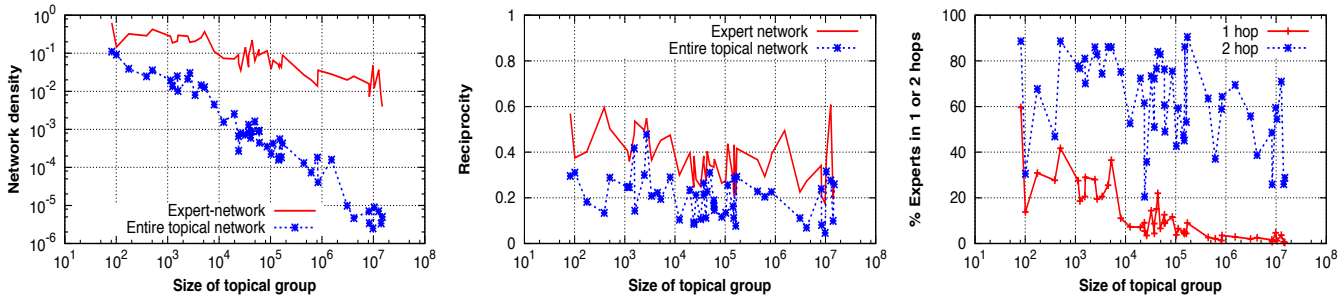


Figure 4: Fraction of tweets in the Twitter 1% random sample, which are posted by experts and seekers in the 50 selected topical groups.



(a) Network density **(b) Reciprocity** **(c) Experts within 1 or 2 hops**

Figure 5: Network properties of the topical-networks (expert-seeker networks) and the expert-networks for the selected topics: (a) network density, (b) reciprocity, (c) mean percentage of all experts who are within 1 or 2 hops from a given expert.

Finally, much of the tweeting activity in Twitter is related to the popular topics. To show this, we considered the 1% random sample of all tweets provided by Twitter, during the month of December 2012, and measured the fraction of tweets that are posted by the experts and seekers in the 50 selected topics shown italicized in Table 3. Figure 4 plots this fraction against the size of the topical groups (measured as the total number of distinct users – experts or seekers – in the groups). The plots show that a considerably smaller fraction of the tweets are related to niche topics (smaller topical groups), as compared to those that are related to popular topics (large topical groups). So studies focusing on popular users or tweets would not capture the diverse set of niche topical groups. However, our analysis reveals that Twitter can also be a source of valuable information on specialized topics.

Following behavior and interactions in topical groups

We now conduct a detailed analysis of the *topical-networks*, i.e., the sub-graph of the Twitter subscription network induced by the set of experts and seekers in a particular topical group. We also separately study the network among the experts within the topical groups, which we refer to as *expert-networks*. Our analysis is done on a set of 50 topical groups that are shown italicized in Table 3 and also indicated in Figure 3.

Component and density analysis

We start by analyzing the number of strongly connected components (SCCs) in the topical-networks and the expert-networks.[2] Across all topics, the expert-network has a single giant SCC, which includes more than 90% of the experts on the topic. The experts who are not within the giant SCC are almost always singletons who are not directly connected to any other expert on the same topic. The entire topical-network also has a single giant SCC which contains almost all experts on the topic. However, a large majority of the seekers (who are not experts themselves) are not included in the giant SCC, which suggests that seekers are only loosely connected with the experts and other seekers. Intuitively, one would not expect seekers to be frequently followed by other seekers or experts.

[2]We also analyzed weakly connected components (WCCs), but the results are less interesting – across all topics, the entire topical-network forms a single WCC. In case of expert-networks as well, more than 99% of experts are in a single WCC for almost all topics.

Next, we computed the network density of the topical-networks and the expert-networks on each topic. We plot the density values in Figure 5(a) against the size of the topical group. Note that all the plots given henceforth have the size of topical groups on the x-axis – this is to highlight the fact that the niche topical groups corresponding to less popular topics (which appear towards the left in the plots) often exhibit very different characteristics than the large groups corresponding to the most popular topics (which appear towards the right in the plots).

Figure 5(a) shows that the topical-networks for the popular topics have very low densities. In fact, for the most popular topics, the densities of the topical-networks are comparable to that of the entire Twitter network (which is of the order of 10^{-7} [16]). However, the densities are relatively much higher for the niche topics.

Figure 5(a) also shows that across all topics, the expert-networks have much higher densities than the entire topical-networks. In particular, the expert-networks in niche topics have very high densities; for instance, for almost all topics having less than 1000 experts, the density of the expert-network is higher than 0.2. This shows that experts, specially in niche topics, are very well connected to other experts on the same topic. These results indicate the presence of strong *homophily* among the experts, in which the establishment of links occurs due to "immutable" similar characteristics [10], like expertise on a common topic. However, seekers do not exhibit strong homophily; they connect to other group members very sparsely, which explains the relatively low densities of the entire topical-networks.

Reciprocity

Next, we study the reciprocity of links in the topical-networks. For a given topic, we measure the fraction of links in the topical-network that are bi-directional, as a fraction of all node-pairs that are linked by at least one link (i.e., fraction of cases where both links $u \rightarrow v$ and $v \rightarrow u$ are present between users u and v). Figure 5(b) plots the reciprocity within the topical-networks and the expert-networks for the selected topics, against the size of the topical group. We observe that across all topics, reciprocity in the entire topical-network is low; for many of the topical-networks, reciprocity is lower than or around 0.22, which was observed to be the reciprocity of the global Twitter network [16]. This low reciprocity is ex-

pected, since experts are not likely to reciprocate links from ordinary seekers. However, reciprocity is relatively higher for the expert-networks, mostly varying between 0.3 and 0.6. Further, reciprocity is generally higher for the niche topics, and reduces for the more popular topics. However, there are a few popular topics that have high reciprocity, such as 'business' (0.6) and 'health' (0.49).

Proximity to experts

Network proximity to experts is important, since the closer a user is to an expert, the higher the chance that the user can receive information tweeted by the expert. Users get all information tweeted by their 1-hop neighbors and they can hear tweets from experts farther away through retweets. But, the chance of hearing information through retweets decreases dramatically with distance. So, next, we measure the fraction of all experts in an expert-network that lie within 1 or 2 network hops from an individual expert.

Figure 5(c) shows, for each of the 50 selected topical groups, the mean percentage of all experts in the group who are within 1 or 2 hops from a given individual expert (the mean is taken over all experts in the topical group). The plots show an interesting trend. While the fraction of all experts within 1-hop is *low*, particularly for the popular topics, the fraction of experts within the 2-hop neighborhood of an individual expert is generally high (60% – 80%) for most topics, including popular topics. Thus, a majority of experts on a topic lie within a 2-hop distance of any given expert on the topic. This is probably because the experts in a topical group are connected to each other relatively densely (see Figure 5(a)); hence, even if one connects to only a few experts in a topical group (1-hop), her 2-hop neighborhood would contain a significant fraction of all the experts in the group.

The above observation has a number of important implications for search and recommendation systems on Twitter. Most experts directly follow only a small fraction of all experts, so they are likely to receive only a fraction of the interesting content being posted by all experts on a topic. This motivates the need for topical search / recommendation systems that can suggest experts and information missing from the 1-hop neighborhood. The fact that a majority of experts could be found within one's 2-hop neighborhood suggests that search / recommendation systems could focus on mining information from the 2-hop neighborhood of a user.

Interactions among experts in topical groups

We now investigate whether the experts in a topical group personally interact with one another. In Twitter, a user u can personally interact with user v through a 'directed message' to v, by including a user-mention "@v" in a tweet. We say an expert u interacts personally with another expert v if u sends at least one directed message to v, i.e., if u mentions v at least once. We collected the tweets posted by all the topical experts during December 2012, and obtained the user-mentions in the tweets. For a given topical group, we construct a *mention-network* among the experts in the group, where there exists a directed edge $u \rightarrow v$ if u mentioned v at least once. Fig. 6 plots the density of the mention-networks for the experts in the 50 selected topical groups, against the group size.

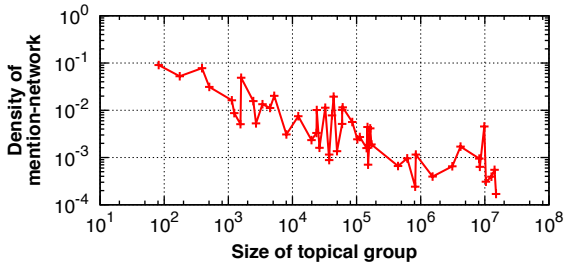


Figure 6: Density of the mention-network among experts for the selected topical groups.

We find that the densities of the mention-networks are much lower than the densities of the corresponding expert-networks (which consider subscription links, shown in Fig. 5(a)). This implies that the experts do *not* personally interact with most of the other experts in the same topical group that they are a part of.

Tweeting behavior in topical groups

In this section, we study the nature of the information being posted by experts in the topical groups. Specifically, we analyze whether experts tend to tweet more frequently on the topics of their expertise, and whether the content posted by multiple experts can be utilized to develop topical recommendation systems.

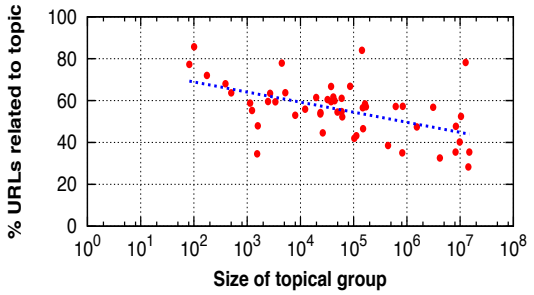
Methodology to ascertain topic of tweets

As stated earlier, we collected all tweets posted by all the topical experts on the 50 selected topics during the month of December 2012. We decided to focus on the tweets containing URLs (similar to [30]) as it is easier to infer topics related to the Web pages pointed to by the URL. We considered a tweet to be related to a topic if the URL contained in the tweet is related to the topic, which we ascertained using the following methodology.

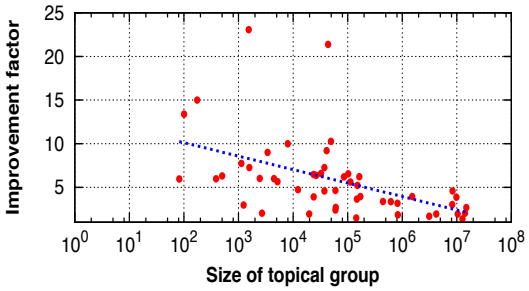
We first obtained a set of keywords related to each topic. For a given topic T, we considered the experts in that topic and checked the 5 most frequent words extracted from the Lists of each expert (as described in the section *Inferring Topical Groups*). We selected the top 10 words which were associated with the most number of experts on topic T as the set of keywords related to T. For instance, for the topic 'religion', the set of keywords consists of words such as, 'christian', 'church', 'faith', 'catholic', 'atheists', 'pastors', and so on. These are essentially the words which co-occur most frequently with the word 'religion' in the Lists of the identified experts on religion.[3]

Next, to ascertain whether a URL is related to a specific topic, we used the AlchemyAPI service (`www.alchemyapi.com`), which uses natural language processing algorithms to extract semantic keywords from the contents of webpages. We considered a URL to be related to a topic if any of the keywords identified by AlchemyAPI for this URL occur within the set

[3]This is similar to the popular Information Retrieval technique of expanding search queries using words which frequently co-occur with the search terms [15].



(a) Fraction of expert-URLs on topic



(b) Expert-URLs vs random-URLs

Figure 7: (a) Percentage of expert-URLs on a topic (randomly sampled from among the URLs posted by experts in the corresponding topical group) which are related to the topic (b) Comparing the percentage of expert-URLs which are related to the topic, to the percentage of random-URLs (obtained from the 1% random sample provided by Twitter) related to a topic.

of keywords related to the topic, obtained by the aforementioned method.

Do experts tweet on their topic of expertise?

For each of the 50 selected topics, we randomly selected a set of 1000 distinct URLs from among the URLs posted by the experts on the topic. We refer to these sets as the *expert-URLs* on a topic.[4] As a baseline for comparison, we also randomly selected 1000 distinct URLs from Twitter's 1% random sample tweet stream. We refer to this set of randomly selected URLs as the *random-URLs*. We then used the AlchemyAPI service to extract topical keywords for each URL in the expert-URLs and the random-URLs sets.

Figure 7(a) plots the percentage of expert-URLs which were found to be related to the corresponding topic, against the size of the corresponding topical group. We find that for many of the groups, a large majority of the expert-URLs (posted by experts in that group) are related to the specific topic. The match is especially high for the niche topical groups, as shown by the decreasing nature of the logarithmic curve of best fit (obtained by the least-squares method) with increasing group size. Specifically, for as many as 36 groups, more than 50% of the expert-URLs were found to be related to the specific topic. These include mostly niche topical groups, but also a few popular groups such as 'books', 'government', 'fashion' and 'environment'.

Expert-URLs vs. random-URLs

We now examine whether expert-URLs (posted by experts in a topical group) tend to be more relevant to the topic of their expertise than random-URLs. We find that for many of the topics, especially the niche ones, the fractions of expert-URLs related to the specific topics are far higher than the fraction of random-URLs which are related to these topics. To quantify this, we measure the ratio between the percentage of expert-URLs and the percentage of random-URLs that were found relevant to a topic; we refer to this ratio as the *improvement factor* for the topic. Figure 7(b) plots the improvement factor for the 50 selected topics, against the size of the corresponding topical groups. It can be seen that the improvement factor is especially high (more than 10) for the niche topics such as 'psychology', 'karate', 'theology', 'neurology' and 'astrology', which are very rarely represented in the Twitter random sample.

[4]Throughout this section, we consider at most 1000 URLs per topic due to rate-limits on the use of the AlchemyAPI service.

Random URLs		Chemistry URLs		Malaria URLs	
Keyword	%URLs	Keyword	%URLs	Keyword	%URLs
likes	16.8	researchers	13.2	malaria	26.4
report	15.6	university	12.8	disease	11.6
gifts	10.8	science	8.0	cases	11.6
language	10.4	scientists	7.2	countries	11.2
gift	10.4	information	7.2	deaths	10.8
logos	10.0	team	6.8	mosquitoes	10.8
buttons	10.0	study	6.8	health	9.6
answers	10.0	results	6.4	africa	9.2
contact	9.2	research	6.0	number	8.8
sign	8.4	company	4.8	study	8.8
click	8.4	work	4.8	treatment	8.8
account	8.4	project	4.0	children	8.8
privacy	8.0	energy	4.0	dec	8.8
notes	8.0	percent	4.0	information	7.2
video	7.6	findings	3.6	areas	6.8

Table 4: Top 15 keywords identified by AlchemyAPI for (i) random-URLs derived from the 1% random sample of tweets provided by Twitter, (ii) expert-URLs in topic 'malaria', (iii) expert-URLs for topic 'chemistry'. Also shown are the percentage of URLs for which a given keyword was identified by AlchemyAPI.

Table 4 illustrates this further by showing the top 15 keywords identified by AlchemyAPI for most of the random-URLs in Twitter, and for the expert-URLs in the niche topics 'chemistry' and 'malaria'. For each keyword, the percentage of URLs (in the corresponding set) for which the given keyword was identified by AlchemyAPI is also given. It is clear that the experts in these topical groups primarily post content related to their specific topics of expertise, which are not captured in the Twitter random sample.

Note that some prior studies have attempted to use the random tweet sample provided by Twitter for topical search or recommendation [7]. Our analysis shows that the content posted by topical experts is a much richer source of topic-specific information than random tweet samples, especially for the niche topics. In the next section, we shall investigate whether the content posted by experts can be used for topic-specific recommendations.

IMPLICATIONS OF OUR FINDINGS

In this section, we discuss the implications of our detailed characterization of topical groups in Twitter. Specifically, we focus on implications for detecting topical groups in Twitter, and in developing improved topical recommendation systems in the future.

Topical groups are largely identity-based groups

As discussed in the Related Work section, several sociological theories have been proposed to explain the formation of groups / communities of users in social networks. According to the well-accepted common identity and common bond theory [22, 23], groups can be identity-based or bond-based, depending on the amount of reciprocity and personal interactions among the members, and the topicality of the discussions among the members [13]. We examined the topical groups for these characteristics, and observed low density and reciprocity and low levels of personal interactions (as estimated by the @user mentions). Also, a large fraction of the content posted by the experts in these groups are related to the specific topic of interest of the members of the groups. Based on these observations, it is clear that the topical groups closely resemble identity-based groups.

We also find that the experts in the topical groups post a significant amount of popular information (e.g., their tweets are highly retweeted by their followers) on the specific topics of interest of the group, hence users interested in a topic have a natural tendency to subscribe to these experts. Thus the formation of these topical groups can be explained by the common identity and common bond theory as being driven by the common interests / expertise of the members on the corresponding topic.

Detectability using community-detection algorithms

Another common approach to identify groups of related users is to use community detection algorithms [11] on the Twitter network graph. Since community detection algorithms rely only on the graph structure, they tend to scale better than semantic approaches that require gathering semantic annotations. However, prior studies [13] have shown that the communities identified by these algorithms are usually closer to bond-based groups and are less likely to detect identity-based groups (like our topical groups). Hence we now investigate whether our topical groups we detected could have been identified by such algorithms.

Using graph theoretic approach to identify communities

To investigate the pros and cons of the graph theoretic approach, we use the well-known BGLL community detection algorithm [5] on the Twitter subscription network described earlier.[5] BGLL detected a total of 686,296 communities. However 375,743 (54%) of these communities are of size 2, i.e., they are the isolated pairs of users only following each other. Figure 8 shows the sizes of the remaining 310,553 BGLL communities. They exhibit a wide variation in their size, with the median and average sizes of the communities being 3 and 108 members, respectively.

[5]Since BGLL runs on undirected networks, we considered all links of the Twitter network to be undirected.

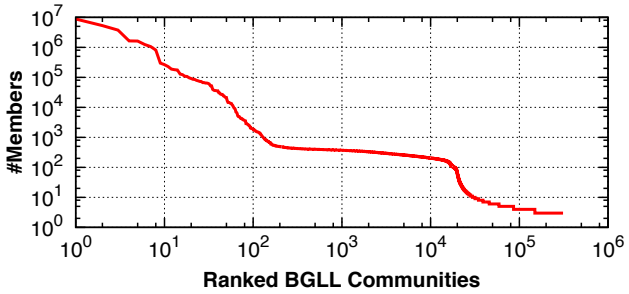


Figure 8: Number of members in the BGLL communities detected in the Twitter network, where communities are ranked in decreasing order of the number of members.

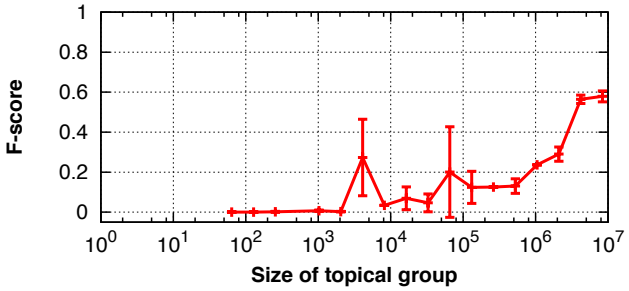


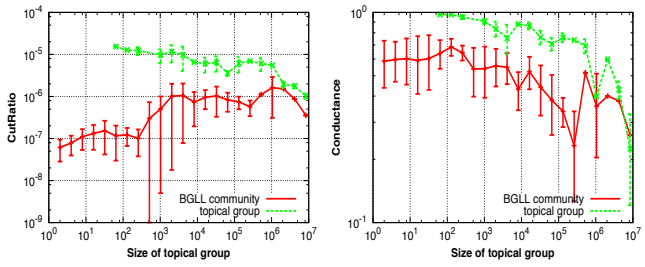
Figure 9: F-Score values between each topical group and the best matched BGLL community. The F-Scores for niche topical groups are low, indicating they are hard to detect by BGLL.

Comparing graph based approach and semantic approach

We begin by checking whether BGLL communities are able to capture at least some of the semantically meaningful topical groups. To this end, we considered the 50 topical groups (detected using the semantic approach) that were shown *italicized* in Table 3, and attempted to find the BGLL communities that most closely match each one.

We found the best matching BGLL community for each of the selected topical groups, according to the F-Score metric [31]. For a given pair of candidate BGLL community and topical group, the F-Score is the harmonic mean of the *precision* and *recall*, where the precision is the fraction of the members of the topical group who are also included in the candidate BGLL community, and the recall is the fraction of the members of the candidate BGLL community who are also included in the topical group. The F-Score takes values between 0 and 1, where 1 represents perfect match while 0 represents no match [31].

Figure 9 shows the F-Score for the selected topical groups, arranged in increasing order of the size of the group (i.e., the number of users in the topical group). It is seen that the smaller or niche topical groups (towards the left in the figure) have very low F-Score with even the best matched BGLL community. This shows that the graph theoretic approach (using BGLL community detection algorithm) is unable to detect the topical groups found using semantic approach, specially the smaller (niche) ones. In fact, we found that for a majority of the semantically meaningful topical groups, their members are spread over more than 100 different BGLL communities.



(a) Cut-ratio (b) Conductance

Figure 10: Comparison of community quality metrics between topical groups and BGLL communities. The cut-ratio and the conductance for topical groups are higher than for BGLL communities, thus making topical groups harder to detect using graph theoretic approaches.

Graph structure of topical groups

One reason why community detection algorithms might fail to detect semantically meaningful topical groups is that the semantic groups do not constitute good-quality communities *structurally*, i.e., the nodes do not form a distinctive community structure in the network graph. To test this hypothesis, we computed several metrics for community quality for the selected topical groups, and also for the BGLL communities. These metrics [18] indicate the quality of the community structure (within the global Twitter network) formed by the members of these topical groups and hint at the potential detectability of these topical groups using graph theoretic approaches. We consider the following two community quality metrics:

(1) *Cut-ratio*: The fraction of the number of external edges leaving a community to the total number of possible external edges. A lower cut-ratio implies better community structure.

(2) *Conductance:* The fraction of the number of internal edges inside the community to the total edges that point outside the community. A lower conductance shows better community structure.

Figure 10 shows the values of these metrics for the selected topical groups, as well as the BGLL communities. The BGLL communities exhibit good community structures (they do well on measures for community quality), while the topical groups show significantly worse community structure than the BGLL communities. These observations suggest that the topical groups do not conform to the standard graph theoretic measures of community quality, making it almost impossible to detect them through standard *graph-theoretic, semantic-agnostic* community-finding methods.

Developing topical recommender systems

We discuss how our improved understanding of topical groups can be leveraged towards (i) discovering more experts (missing members) in different topical groups, and (ii) discovering important topical information of interest to a topical group.

Recommending topical experts

Given a set of (known) experts in a topical group (i.e., on a topic), we investigate whether the expected tight interconnectivity between all experts in the group could be used to discover and recommend new (yet unknown) experts on the corresponding topic. Intuitively, the yet to be discovered experts would also be well connected to the known experts. So a simple technique to discover new topical experts would be to identify users in the Twitter network that are followed by the highest number of known experts and check their expertise. Such a technique might also be quite efficient as the number of Twitter users followed by multiple experts in a topical group tends to be quite small, especially for niche topics.

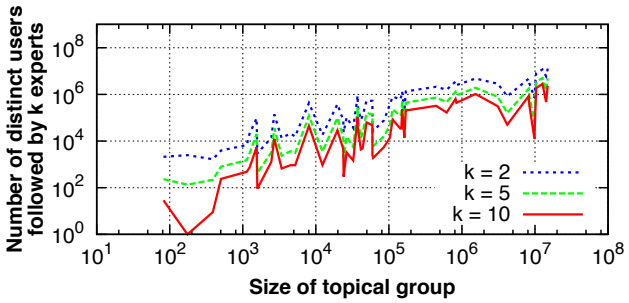


Figure 11: Number of distinct users who are followed by k known experts in a topical group, shown for k = 2, 5, 10.

Topical group	# known experts	# new experts discovered	
		on same topic	on any topic
Malaria	19	17	460
Karate	20	20	348
Chemistry	90	96	1342
Geography	124	103	2615
Sociology	157	231	4450
Physics	255	163	6673

Table 5: Number of new experts discovered among the users followed by known experts.

Figure 11 plots the number of distinct users who are followed by k experts in a topical group (for various values of k), against the group size. It shows that the number of distinct users followed by (k = 10) known experts is of the order of few hundreds for the niche topical groups to few million in the case of the most popular topical groups. We tested the likelihood that many of these users are yet undiscovered experts within their corresponding groups. We checked the expertise of these users using the same methodology as in [12, 25]. Table 5 shows the number of new experts that we discovered in the process for some niche topical groups. For several of the groups, the number of newly discovered experts doubles the size of known experts. These newly discovered experts are outside of the 38.4 million user accounts that we crawled as the seed data for our study.

These results show that an effective and efficient technique to search for new topical experts would be to look for users that are followed by a large number of known Twitter experts. Such a technique could not only be used to keep the list of experts up-to-date as new experts join Twitter in increasingly large numbers every day, but it could also be used to recommend other relevant topical experts to knowledge seekers who are already following a small set of known experts.

Recommending topical content

There have been several efforts to recommend interesting topical content to Twitter users who are interested in certain top-

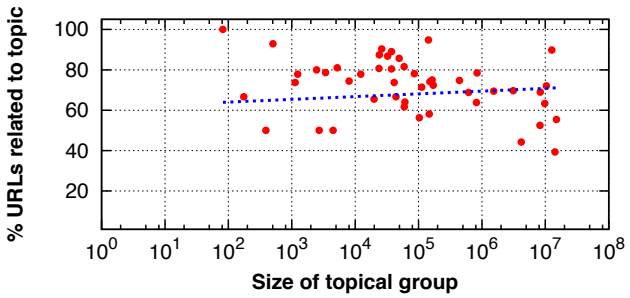


Figure 12: Fraction of the URLs posted by at least 3 distinct experts in a topical group, that are related to the topic. The results show that URLs that are tweeted by multiple experts in a topical group can be used for topical recommendations.

ics [6, 7]. A crucial challenge for such systems is to discover interesting content on specific topics. We now investigate whether the knowledge of a large set of topical experts can be utilized for recommendation of topical content. Intuitively, content that is posted by multiple experts in a topical group is more likely to be relevant and authoritative on the corresponding topic. Hence, for each topical group, we selected the URLs which were posted by at least 3 distinct experts in the topical groups, and used the AlchemyAPI service to check what fraction of these URLs were related to the topic.The results are plotted in Fig. 12, against the size of the topical groups. For 39 out of the 47 selected groups[6] , more than 60% of the expert-URLs that are posted by 3 or more experts were related to the specific topic. Interestingly, this is true not only for niche topical groups (e.g., 'malaria', 'iphone', 'geology', 'linux', 'medicine' and 'physics') but also for popular groups such as 'business', 'sports', 'health' and 'fashion'.

The above result shows that the content posted by multiple experts within a topical group is particularly useful for generating content recommendations to the users interested in that specific topic. Such recommendations are particularly useful for niche topics, where existing search / recommendation systems that rely on random samples from Twitter would struggle to provide good recommendations [7].

CONCLUSION

The primary contribution of this paper lies in identifying topical groups in Twitter at scale, and analyzing the characteristics of the topical groups. We utilized crowd-sourced social annotations to infer the topics of expertise and interests of individual Twitter users, and hence identified topical groups consisting of experts and seekers of information on specific topics. The identified topical groups span a wide diversity of topics, ranging from the most popular topics (e.g., 'politics', 'tech', 'music') to niche, specialized topics such as 'geology', 'forensics', and 'neurology'. We also show that these topical groups exhibit characteristics of identity-based groups, which makes it extremely difficult to detect these groups using graph theoretic community detection algorithms on the Twitter network graph.

[6]For three of the niche topical groups having only a few tens of experts, no URL had been posted by 3 distinct experts, hence this analysis was conducted for the remaining 47 groups.

This study firmly establishes that the popular Twitter microblogging site, beyond providing a platform for celebrities and news media to disseminate information about topics of general interest, is also facilitating the self-organization of a large set of niche topical groups consisting of local, topic-specific experts and dedicated followers. Finally, our analysis of network connectivity and tweeting behavior of users in such topical groups reveal several insights that have implications for designing search and recommendation services on Twitter. For instance, we show that the existing experts can act as a seed set to discover new experts, thus, facilitating dynamic update of the topical group. We also show that content posted by multiple experts in a topical group has a high probability of being topically relevant – this observation can be leveraged to generate topically relevant tweet recommendations. We believe that an interesting direction of future research would be to exploit the concept of topical groups for building such new services.

ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers whose suggestions greatly helped to improve the paper. This research was supported in part by a grant from the Indo-German Max Planck Centre for Computer Science (IMPECS). P. Bhattacharya was supported by a fellowship grant from Tata Consultancy Services.

REFERENCES

1. Twitter Statistics, May 2013. **http://www.statisticbrain.com/twitter-statistics/** (Accessed October 15, 2013).
2. An, J., Cha, M., Gummadi, K., and Crowcroft, J. Media landscape in Twitter: A world of new conventions and political diversity. In *Proc. AAAI ICWSM* (2011).
3. Ardon, S., Bagchi, A., Mahanti, A., et al. Spatio-temporal analysis of topic popularity in Twitter. *arXiv:1111.2904v2 [cs.SI]* (2011).
4. Blanchard, A. L., and Markus, M. L. The experienced "sense" of a virtual community: characteristics and processes. *SIGMIS Database 35*, 1 (Feb. 2004), 64–79.
5. Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 18 (October 2008).
6. Chen, J., Nairn, R., Nelson, L., Bernstein, M., and Chi, E. Short and tweet: experiments on recommending content from information streams. In *Proc. ACM CHI* (2010).
7. Choudhury, M. D., Counts, S., and Czerwinski, M. Find Me the Right Content! Diversity-Based Sampling of Social Media Spaces for Topic-Centric Search. In *Proc. AAAI ICWSM* (2011).
8. Cox, A. What are communities of practice? a comparative review of four seminal works. *Journal of Information Science 31*, 6 (2005), 527–540.
9. De Choudhury, M., Diakopoulos, N., and Naaman, M. Unfolding the event landscape on twitter: classification and exploration of user categories. In *Proc. ACM CSCW* (2012).
10. Easley, D., and Kleinberg, J. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.
11. Fortunato, S. Community detection in graphs. *Physics Reports 486*, 3-5 (2010), 75–174.
12. Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N., and Gummadi, K. Cognos: crowdsourcing search for topic experts in microblogs. In *Proc. ACM SIGIR* (2012).
13. Grabowicz, P. A., Aiello, L. M., Eguiluz, V. M., and Jaimes, A. Distinguishing topical and social groups based on common identity and bond theory. In *Proc. ACM WSDM* (2013).

14. Java, A., Song, X., Finin, T., and Tseng, B. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proc. WEBKDD / SNA-KDD Workshop* (2007).
15. Jones, K. S. *Automatic keyword classification for information retrieval*. Butterworth, 1971.
16. Kwak, H., Lee, C., Park, H., and Moon, S. What is Twitter, a social network or a news media? In *Proc. ACM WWW* (2010).
17. Lerman, K., and Ghosh, R. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In *Proc. AAAI ICWSM* (2010).
18. Leskovec, J., Lang, K. J., and Mahoney, M. Empirical comparison of algorithms for network community detection. In *Proc. ACM WWW* (2010).
19. Lin, C. X., Mei, Q., Han, J., Jiang, Y., and Danilevsky, M. The joint inference of topic diffusion and evolution in social communities. In *Proc. IEEE ICDM* (2011).
20. Twitter help center: How to use twitter lists. **`http://tinyurl.com/lists-howtouse`** (Accessed October 15, 2013).
21. McMillan, D., and Chavis, D. Sense of community: A definition and theory. *Journal of Community Psychology 14*, 1 (1986), 6–23.
22. Prentice, D. A., Miller, D. T., and Lightdale, J. R. Asymmetries in attachments to groups and to their members: Distinguishing between common-identity and common-bond groups. *Personality and Social Psychology Bulletin 20*, 5 (1994), 484–493.
23. Ren, Y., Kraut, R., and Kiesler, S. Applying Common Identity and Bond Theory to Design of Online Communities. *Organization Studies 28*, 3 (2007), 377–408.
24. Sassenberg, K. Common bond and common identity groups on the Internet: Attachment and normative behavior in on-topic and off-topic chats. *Group Dynamics Theory Research And Practice 6*, 1 (2002), 27–37.
25. Sharma, N., Ghosh, S., Benevenuto, F., Ganguly, N., and Gummadi, K. Inferring Who-is-Who in the Twitter Social Network. In *Proc. WOSN Workshop* (2012).
26. Talbot, D. How Google Ranks Tweets, Jan 2010. **`http://www.technologyreview.in/web/24353/`** (Accessed October 15, 2013).
27. Wagner, C., Liao, V., Pirolli, P., Nelson, L., and Strohmaier, M. It's not in their tweets: Modeling topical expertise of twitter users. In *Proc. ASE/IEEE SocialCom* (2012).
28. Wang, L., Lou, T., Tang, J., and Hopcroft, J. E. Detecting community kernels in large social networks. In *Proc. IEEE ICDM* (2011).
29. Wenger, E. *Communities of practice: Learning, meaning, and identity*. Cambridge University Press, New York, NY, USA, 1999.
30. Wu, S., Hofman, J. M., Mason, W. A., and Watts, D. J. Who says what to whom on Twitter. In *Proc. ACM WWW* (2011).
31. Yang, J., and Leskovec, J. Defining and evaluating network communities based on ground-truth. In *Proc. MDS Workshop* (2012).