Characterizing Twitter Elite Communities:Measurement, Characterization, and Implications

Reza Motamedi, Saed Rezayi, Reza Rejaie, Ryan Light
University of Oregon
{motamedi,saed,reza,light}@uoregon.edu

Walter Willinger
Niksun, Inc.
wwillinger@niksun.com

Keywords

Social Networks, Connectivity Structure, Social Context

ABSTRACT

This paper presents a new, socially-informed approach for characterizing the connectivity structure among Twitter users. We primarily focus on a sub-graph of top 10K most-followed users (or elites) that we refer to as *elite network*. We present a new technique for efficiently capturing the Twitter elite network along with social attributes of individual elite nodes. We show that the elite network (even at smaller sizes) is composed of a 15-20 resilient elite communities that all exhibit a clear social cohesion. These characteristics imply that the elite communities represent "socially meaningful" components of the Twitter structure and offer a coarse view of the Twitter elite network.

We then characterize the community-level structure of the elite network and identify the pairwise tendencies between elite communities to follow each other. We also assess the cross-influence between elite communities based on retweeting and replying and show that such influences are effectively contained within individual elite communities. Finally, we illustrate that most regular (non-elite) Twitter users tend to primarily follow (*i.e.*, show interest to) users in a single elite community. A group of regular users who primarily follow an elite community form its "shadow partition". We show that the fraction of relationships between elites that span across elite communities is very similar to the fraction of relationship between regular users that span across different shadow partitions. This suggests that elite communities sketch a socially-aware and coarse view of the entire Twitter structure.

1. INTRODUCTION

The increasing popularity of online social media (*e.g.*, Twitter) in recent years has fueled the growing interest in understanding their connectivity structure and the exchange of information as well as influence among their users. A rather common data-driven approach for characterizing the connectivity structure of these networks is to identify their important components (*e.g.*, tightly connected group of nodes such as communities [3], or regions [15]) and represent the structure as a collection of inter-connected components. Such a coarse view is often much smaller and less complex and thus reveals the main connectivity features of the structure more clearly. However, the number of such components could still be very large (*e.g.*, Twitter consists of 24K communities). More importantly, the usual absence of any social (or other) context for individual components in these studies makes it difficult to interpret their role and properly assess their importance in the overall structure.

There is a wealth of established sociological theory on the structure of non-virtual social networks. While these theories and concepts offer valuable insights, their relevance and applicability to the structure of online social media is still being examined [9, 19]. On the one hand, these online spaces empower individual users to connect and interact across cultural and geographic boundaries. This suggests that usual hierarchies in offline structures may not be as pronounced in online social media [17]. On the other hand, there is evidence of the reproduction of offline groups in online spaces [14]

In particular, the concept of *imagined communities* [25] captures the strong sense of community a group, such as the citizens of a nation, may feel even without face-to-face communication based on cultural or contextual overlap. This concept is particularly relevant to online social media, such as Twitter, because it is built around both asymmetrical broadcasting characteristic of traditional media and reciprocal sharing characteristic of many social networks. This in turn raises the following important questions: *(i) Whether the structure of an online social media is composed of such imagined communities*, *(ii) Whether these communities affect the interactions and influence among users?* and *(iii) How such communities can even be identified in the huge and complex structure of a popular online social media?*

In this paper, we tackle these important questions in the context of Twitter. Our first contribution is our proposed methodology to identify socially meaningful components in Twitter. To this end, we primarily focus on the subgraph that connects the top 10K most-followed Twitter users. We refer to these users as "Twitter elites" and the resulting subgraph as "Twitter elite network". Such an elite network represents the "backbone" of the Twitter structure as the elites collectively reach more than 80% of all Twitter users. Furthermore, the individual elite users often exhibit a clear social context that can be captured while such information is unclear or unavailable for regular users. We present a new technique to efficiently capture the subgraph of highest-degree nodes in a large graph as our second contribution. Using our technique, we capture and validate the top 10K Twitter elite network, and then collect the social and geographic attributes of elite users from online sources. We consider the elite network at different sizes (less than 10K), and at each size detect its resilient communities that we call *elite communities*. We examine the social cohesion of elite communities as well as their inter-connectivity structure and cross influence based on retweet and reply. Finally, we explore whether elite communities can offer any insight about the grouping of regular Twitter users.

Key Findings: The third contribution of this work is a collection of insightful findings that can be summarized as follows: First, we show that the Twitter elite network at various sizes consists of 15-20 modularity-based elite communities that clearly exhibit cohesion around a specific social, geographic or more subtle theme. Furthermore, as we expand the size of the elite network, individual communities grow, merge, or split but the collection of their high level themes remain rather stable. These findings indicate that Twitter elite communities represent socially meaningful components of the Twitter network. Second, we examine the community-level structure of the elite network and characterize any bias in the connectivity between elite communities using two different measures. These analysis reveals a tighter connectivity between a small subset of elite communities and examines the role of specific nodes that act as bridge between these communities. Third, we assess the cross influence between elite communities based on retweeting and replying and show that such a cross influence among elite users (based on either measure) is contained within individual elite communities. Finally, we illustrate that a majority of the elite friends of regular Twitter users tend to be in a single elite communities. This in turn offers a promising criteria to group regular users into "shadow partitions" based on their association with elite communities. We shows that the level of overall inter-connectivity between shadow partitions mirrors the same characteristics for the elite communities. This suggests that the shadow partitions can be viewed as the extension of their corresponding elite community. In other words, elite communities offer a coarse view of the entire twitter structure.

The rest of the paper is organized as follows: In Section 2, we present our technique for capturing the Twitter elite network. Our approach for detecting elite communities and their basic characteristics are descried in Section 3. The inter-connectivity and cross influence among elite communities are discussed in Section 4 and 5, respectively. We describe the association of regular users with individual elite communities and how this could be leveraged to group the users into shadow partitions in Section 6. We conclude the paper and present our future plans in Section 7.

2. CAPTURING ELITE NETWORK

Our goal is to efficiently capture the Twitter elite network - that is a subgraph of Twitter that includes the top-N most-followed accounts (*i.e.*, node) and the friend-follower relationships among them (*i.e.*, edges)[1]. Furthermore, we need to annotate each node with its social and geographical (location) attributes for our analysis.

Our data collection strategy for capturing Twitter elite network consists of the following four steps: *(i)* Capturing a list of most-followed Twitter accounts through public resources and random walks used as seeds. *(ii)* Inferring their pairwise connections. *(iii)* Identifying missing accounts, validating the information, and collecting pairwise connections. *(iv)* Collecting all profile information and available tweets of qualified accounts. The details of individual steps are as follows:

Step 1: To bootstrap the data collection process, we crawl lists of the most followed accounts from online resources. In particular, marketing websites such as socialbakers.com offer professionally maintained lists of most followed accounts in variety of OSNs in different social categories (*e.g.*, celebrities, actors, sport, community, ...). Each list on socialbakers.com provides up to 1000 top accounts in the selected category along with the number of followers and username for each account. We collect the list associated with all offered categories and subcategories and create a unified list that includes all the uniquely-discovered user accounts with their number of followers (and associated rank), their category and location. This resulting unified list consists of 59 832 unique users whose number of followers varies from 263 to 81M, and they are associated with 123 categories and 191 unique countries.

We also conduct approximately 2K random walks on the list of friends from randomly selected Twitter accounts to identify high-degree nodes. These random walks are biased towards users with more followers and offer an efficient technique to identify users with most followers [24, 20]. Equipped with these two techniques to identify potential highest degree nodes, we then create a master list that includes more than 60K accounts. We mainly focus on the top 10K accounts with the most followers from this master list. In this list, 89% are exclusively reported on socialbakers.com, 3.2% are exclusively identified through random walks, and 7.8% are found through both techniques. It is worth noting that the overall popularity rank of the accounts exclusively found by random walk is at least 133 out of 10K.

Step 2: It is prohibitively expensive to find all the pairwise connections between the identified accounts by collecting and examining all their followers. Our key observation is that the number of friends for elites are almost always several orders of magnitude smaller than the number of followers. Therefore, instead of followers, we collect the complete list of friends for each selected account from Twitter (using its API). This implies that the connection between account u_{fri} and its follower account u_{fol} (denoted as $u_{\text{fri}} \rightarrow u_{\text{fol}}$) is discovered when we collect the friend list of account u_{fol}, *i.e.*, each edge is discovered from the follower side. This crawling strategy significantly reduce the overhead of capturing all links between identified accounts. The total number of crawled friend-follower relationships with this strategy is 504.8M which consists of 95M unique friends for the top 10K most-followed elites.

Step 3: At this point, we have a snapshot of the most-followed Twitter accounts and their pairwise directed connections. It is indeed possible that the identified top 10K accounts so far do not accurately capture the top 10K accounts on Twitter, *i.e.*, some elite accounts might be missing. We take a few steps to verify whether the collected information is correct and complete. Our final step is similar to the approach proposed by Avrachenkov *et al.* [1]. The observation is that any such missing elite account should be followed by many elites already identified as top 10K accounts. Note that we already obtained the entire list of friends for top 10K accounts. We calculate the number of elite-followers for all these collected friends that are not among the elites, and sort the resulting list by the number of elite-followers. We start by scanning this list from the top and collect account information including the number of followers for users in this list. If the number of followers for any of these accounts is larger than the number of followers for the account at rank 10K, we add it to the master list (at the proper rank) and update the ranks for all elites. We continue this process until 100 consecutive accounts from this sorted list do not make it to the master list. We finally identify the edges between these newly added accounts and other top 10K accounts by collecting their friend list. Using this technique, we detected 264 (2.6%) missing accounts that are between the rank of 500 and 10K. The small percentage of missing accounts along with their relatively low ranking indicate that our master list is accurate. All in all, among the top 10K most followed accounts, 8 704 were exclusively reported in socialbakers.com, 301 were found exclusively using random walks, 731 are from both the mentioned resources. Finally, checking the most followed friends of elites placed and 264 are among the friends of most-followed accounts.

Step 4: We collect all the available tweets for the top 10K Twitter

[1] We use the terms *nodes with highest degree* and *most followed accounts* interchangeably.

accounts. The available tweets[2] for each account are used to investigate the influence between elites and gain some insight on how they use Twitter.

Who is Elite? It is certainly compelling to consider Twitter users with the highest number of followers as Twitter elites. One remaining question is *how many most-followed accounts should be considered for forming the elite network?* We argue that the 10K-ELITE offers a sufficiently large view of the elite network in Twitter for several reasons as follows: First, the skewed distribution of the number of followers implies that the number of followers rapidly drops with rank. For example, the top 10 most followed accounts have between 51.9M to 81.7M followers while the last 10 accounts in the top 10K have around 0.4M followers and the median number of followers among the top 10K is 0.8M. Therefore, the popularity (and thus importance) of any account beyond top 10K would be much less.

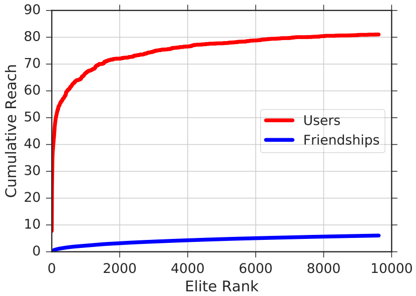


Figure 1: The total number of nodes and edges that are reached by the top-n elite nodes.

Second, examination of the friend list for 10K random twitter users shows that 80% of all twitter accounts follow the top 10K elites. To this end, we collect an unbiased set of random twitter users using random walk based techniques described in [24]. Figure 1 presents the fraction of random users that are direct followers of the top-n elite users as the elite network is extended. As the figure shows, 80% of the random users are immediate followers of the top 10K elites. The figure also shows that the gain from extending the elite network dramatically diminishes as we pass the 2K-ELITE mark. Third, while it is feasible to capture a larger elite network beyond 10K, reliably collecting the desired attributes (social and location) for these users is very expensive and their addition has diminishing return.

To examine whether and how the size of the resulting elite network affects its structural properties, we consider the Twitter elite network at different sizes (or views). Each view, which we refer to as nK-ELITE, contains the top *n-thousand* most-followed accounts and friend-follower relationships between them.

2.1 Overall Structure of the Elite Network

Before we conduct any analysis on the Twitter elite network, we present a number of basic characteristics for each view of the elite network in Table 1, including the number of nodes and directed edges ($|E|$), reciprocity (Rcp), transitivity or clustering coefficient (Tran), and diameter (Diam). We also include the number of connected components and strongly connected components. This table clearly shows that as the size of the elite network is extended (from 1K to 10K), it becomes denser (average degree increases from 49 to 152), the fraction of reciprocated edges initially drops and then increases, and its diameter slightly increases. In all views, 32-40% of the friend-follower relationships are reciprocal, which is higher compared to the reported 22% for the entire Twitter social graph [12]. Interestingly, we observe that all views of the elite network have a single weakly connected component that includes an absolute majority of all nodes except for one or two nodes. However, the number of strongly connected components (SCC) grows roughly proportional with the size of the elite network. The rank correlation between the number of public vs. elite followers for top-10K elite is around 0.55 while the rank correlation between their public vs. elite friends is 0.1, *i.e.*, the popularity of elites among all users and elites are moderately correlated.

Strongly Component Analysis: We conduct (strongly) connected component analysis [10] on different views of the elite networks in order to reveal their overall topological structure. As we reported in Table 1, each view of the elite network has many strongly connected components (SCC). However, the largest strongly connected component (LSCC) in each view contains an absolute majority of all elites while all other SCCs have a single node (and in a few cases a handful of nodes). The right section of Table 1 summarizes the fraction of nodes and edges that are within the LSCC in each view. This table shows that the LSCC in each view contains 91-94% of all nodes and 94-97% of all edges of the corresponding elite network.

To gain more insight into the structure of the elite network, Figure 2 visualizes the strongly connected component structure of 1K-ELITE, 5K-ELITE, and 10K-ELITE as directed graphs where each circle represents a SCC with the number indicating the number of nodes in that SCC. LSCC is shown with a green circle in the center. Arrows represent friend→follower relationships between users in different SCCs. These figures clearly illustrate that in all views the SCCs form a "star-like" structure where the LSCC is in the center and there are a number of directed edges from every other SCC (that we call "outsider") to nodes in LSCC. We recall the direction of edges are from a friend to a follower (or the direction of tweet propagation.) Therefore, Figure 2 indicates that nodes in the LSCC have an interest in and receive tweets from nodes in other SCCs (through the elite network) but the opposite is not true. In fact, more than 99% of outsiders are followed by users in the LSCC. Most outsider nodes are in a single node SCC and few of them consist of two or more nodes. For example, the Pope has four accounts that only follow each other but they are followed by many accounts inside the LSCC.

Basic Characteristics of Elite Networks: As the size of the elite network is extended (from 1K to 10K), it becomes denser (average degree increases from 49 to 152), the fraction of reciprocated edges varies is around 32-40%, which is higher compared to the reported 22% for the entire Twitter social graph [12]. Interestingly, we observe that all views of the elite network have a single weakly connected component that contains more than 99.99% of all nodes. Furthermore, the largest strongly connected component (LSCC) [10] in each view contains 91-95% of nodes and 94-97% of all edges in the elite network. Figure 2 visualizes the strongly connected component structure of 1K-ELITE, 5K-ELITE, and 10K-ELITE as directed graphs where each circle represents a SCC with the number indicating the number of nodes in that SCC. LSCC is shown with a green circle in the center. Arrows represent friend→follower relationships between users in different SCCs. These figures clearly illustrate that in all views, the SCCs form a "star-like" structure where the LSCC is in the center and there are a number of directed edges from every other SCC (that we call "outsider") to nodes in LSCC. We recall the direction of edges are from a friend to a follower (or the direction of tweet propaga-

[2] Twitter only provides the last 3 200 generated tweets by each user.

Table 1: Basic characteristics of the elite networks and their weakly and strongly connected components

					CC			SCC		
View	$\|E\|$	Rcp	Tran	Diam	#CC	$\%\|V\|$	$\%\|E\|$	#SCC	$\%\|V\|$	$\%\|E\|$
1K-Elite	49K	0.35	0.3	7	1	100.0	100.0	64	93.5	94.6
2K-Elite	126K	0.34	0.24	7	2	100.0	100.0	110	94.2	95.6
3K-Elite	231K	0.32	0.2	7	3	99.9	100.0	171	94.1	95.8
4K-Elite	344K	0.31	0.18	7	3	100.0	100.0	231	94.0	95.9
5K-Elite	491K	0.32	0.17	7	3	100.0	100.0	279	94.2	96.1
6K-Elite	648K	0.33	0.16	7	2	100.0	100.0	337	94.1	96.2
7K-Elite	816K	0.34	0.16	7	2	100.0	100.0	370	94.5	96.4
8K-Elite	1.0M	0.37	0.17	7	2	100.0	100.0	401	94.8	96.7
9K-Elite	1.2M	0.4	0.18	8	2	100.0	100.0	439	94.9	96.9
10K-Elite	1.4M	0.42	0.19	9	2	100.0	100.0	454	91.5	97.0

tion.) Therefore, Figure 2 indicates that nodes in the LSCC have an interest in and receive tweets from nodes in other SCCs (through the elite network) but the opposite is not true. In fact, more than 99% of outsiders are followed by users in the LSCC. Most outsider nodes are in a single node SCC and few of them consist of two or more nodes. For example, the Pope has four accounts that only follow each other but they are followed by many accounts inside the LSCC. The rank correlation between the number of public vs. elite followers for top-10K elite is around 0.55 while the rank correlation between their public vs. elite friends is 0.1, *i.e.*, the popularity of elites among all users and elites are moderately correlated.

As more nodes are included in the elite network, other SCCs in one view may be pulled into the LSCC in the next view since the extended view may include more shortcuts. Figure 3 illustrates via Sankey diagram [26] how the LSCC and the outsider in each view are mapped/split to the LSCC and the outsider in the next view[3]. In this figure individual views of the elite network are shown along the x axis. For each view, the two vertical boxes represent LSCC and the outsider. The vertical box at the bottom of each column represents the LSCC and the box on the top represents the outsider. Groups of elites ranked by their number of followers are all presented in the first column alongside 1K-Elite. Extending the elite network adds one of the groups to a view to create the next view, for instance accounts with rank [1K..2K] join 1K-Elite to create 2K-Elite. As the plot shows, more than 95% of these newly added elites join the LSCC and the rest join the other SCCs. An examination of these views also reveals that roughly 13-20% of nodes in other SCCs are pulled into the LSCC in the next view. Note however that a group of other SCCs have no friends (*i.e.*, no incoming edges) and thus remain outside the LSCC regardless of the size of the elite network.

Label	Size	Dens.	Cond.	Theme
US/Pop	2.9K	384	0.26	US celebs/actor/music
Spanish	1.9K	208	0.35	Spanish Speaking
US/Corp	1.3K	242	0.58	US Corporate/Media
Arabic	1K	698	0.13	Arabic Speaking
ID	533	93	0.34	Indonesian
BR	508	162	0.38	Brazilian
PH	475	210	0.46	Filipino
IN	335	185	0.57	Indian
TR	271	87	0.34	Turkish
Unstable	155	268	0.98	Unstable nodes
K-PoP	150	51	0.44	Korean Popstars
TH	28	34	0.63	Thai
Adult	20	57	0.48	Adult/Porn
US/TV	19	541	0.99	US TV channels
GLB/Fun	13	119	0.98	Global Entertainment

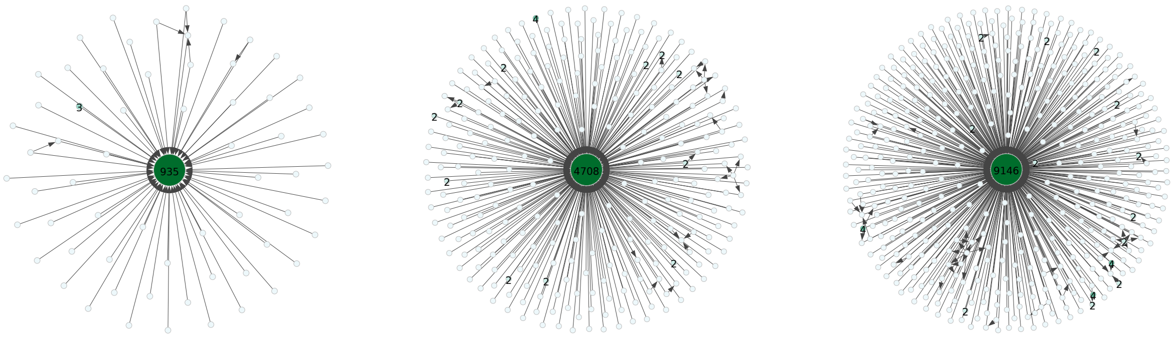
Table 2: Label and key features of 14 elite communities in the Twitter elite network

3. ELITE COMMUNITIES

Our goal is to determine whether the elite network is composed of a collection of meaningful components. The most natural components are groups of tightly connected nodes (or communities). We need to address two issues before applying community detection: First, most commonly-used community detection techniques take undirected graphs as input while the elite network is a directed graph [11]. To address this issue, we first convert each view of the elite network into an undirected graph by converting *each* directed edge into a single undirected edge with the weight of 2 when reciprocal directed edges exist. This representation allows us to encode tighter binds between users with reciprocal edges compare to prior studies (*e.g.*, [13]) where they simply consider a directed graph as undirected. Second, the outcome of the most commonly used community detection techniques (*e.g.*, Louvain [3], BigCalmm [27], InfoMap [22]) is non-deterministic and varies across multiple runs. To address this issue, we use Combo community detection [23] that relies on multi-objective optimization and detects more stable communities across different runs. We also eliminate the residual instability by only considering a group of nodes as a community if they consistently mapped to the same community across different runs. Toward this end, we adopt the following strategy: We run Combo on each view of the elite network k times and determine the communities that individual nodes are mapped to in each run in a vector with k values, called the "community vector". Then, we group all the nodes that are consistently (*i.e.*, all k times) mapped to the same community (*i.e.*, have the same community vector) and refer to the group as a *Resilient Community*. The process of detecting communities also results in group of nodes for which no other node has the same community vector. We group this set of nodes and nodes in resilient communities that are smaller than 10 nodes and refer to them as *Unstable nodes*.

Clearly increasing k is more restrictive and may lead to smaller resilient communities since more runs can simply split a community to two (or more) smaller ones. Figure 4 shows the effect of k on the number of resilient communities identified in 1K-Elite, 5K-Elite and 10K-Elite. As the figure shows, increasing k can split

[3] An interactive visualization of this diagram is available on our project page ix.cs.uoregon.edu/~motamedi/research/elite/evol/sankey_in_out_lscc.html



(a) 1K-ELITE (b) 5K-ELITE (c) 10K-ELITE

Figure 2: The connectivity of strongly connected components of the elite networks.

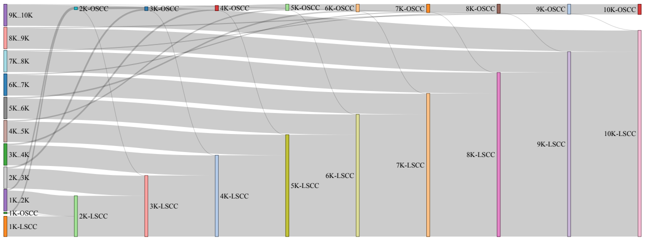


Figure 3: The dynamics of LSCC as the network expands

resilient communities and increase the number of resilient communities. This number may shrink, however, when the newly created resilient communities include less than the minimum threshold size of resilient communities, which we set to 10. Note that in all runs of COMBO, a community smaller than 20 nodes was never identified, *i.e.*, our threshold does not dissolve a community in the *unstable* group. It is also interesting to note that the effect of increasing k is more considerable in 5K-ELITE. This indeed suggests that this view has a less pronounced community-level structure since each run leads to the identification of a very different grouping of nodes as communities [6]. Figure 4 also shows that in all cases the number of resilient communities stabilizes after the initial increase. We conservatively consider $k = 100$ in our analysis, as having more runs does not lead to the identification of more resilient communities in the elite networks.

Table 3 presents the general statistics of the communities identified in each view. As the table shows, COMBO detects between 6-11 regular communities in different runs of various views but the number of resilient communities with 10+ users varies between 10-29 across different views of the elite network and they collectively cover 92-99% of all nodes in each view. Thus, less than 8% of the elites are *unstable* nodes.We emphasize that the identified elite communities are very different from communities on the entire Twitter social graph that contain many regular (*i.e.*, non-elite) users.

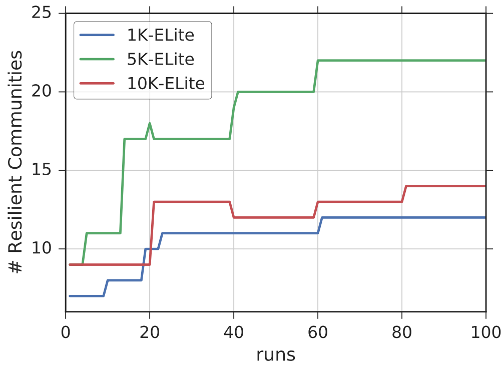


Figure 4: The number of identified resilient communities as a function of k in three views of the elite network.

3.0.1 Resilient vs. Regular Communities

In this subsection, we examine whether the resilient communities exhibit different connectivity characteristics compare to regular communities that are identified by COMBO. This in turn could affect the result of our community-based analysis. We use conductance [4] and modularity [18] as two measure of a graph structure with respect to the identified communities in the graph. Conductance measures how well a certain bipartition of nodes splits in the graph. Therefore, for each community – a cut through the edges in the graph – we can compute a single conductance value. Small conductance values mean that a small number of edges are cut to split the graph into two halves (*i.e.*, the community and the rest of the graph). On the other hand, modularity measures how well a graph divides into modules. In other words, a graph with high modularity computed for a certain grouping of nodes into modules (communities) has dense connections between the nodes within modules, but sparse connections between nodes in different modules. For each graph partitioning into communities a single modularity is computed.

We separately identified communities in each view of the elite network. Figure 5 shows the scatter plot of conductance and size of regular communities identified in all 100 runs of COMBO, the resilient communities, and also the *unstables* in the 10K-ELITE view. We recall that smaller conductance suggests a better separation of the community from the rest of the graph. A close comparison of the communities with the identified regular communities shows that for similar sizes, their conductance values are indeed smaller or similar. There are only two rather small resilient communities that higher conductance compared to regular communities. Also, a very small group of *unstables* have a very high conductance, which

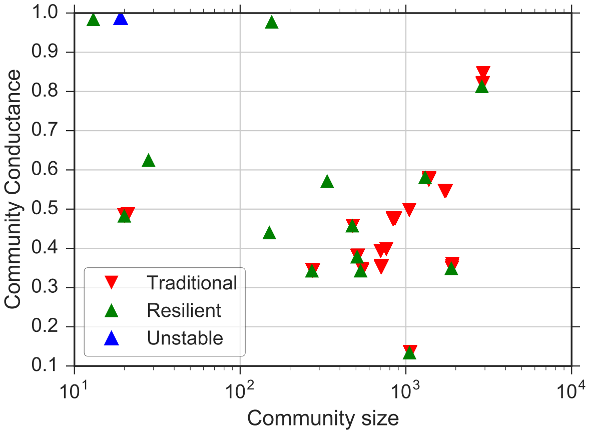


Figure 5: Conductance vs. size of regular and resilient communities and *unstables* identified in 10**K-**E**LITE.**

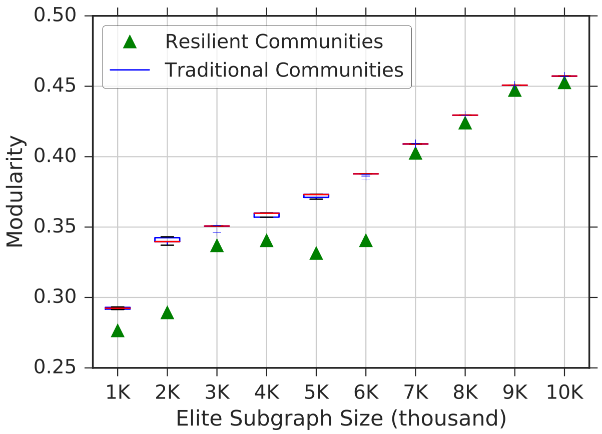


Figure 6: Modulairty of resilient communities and the summary distribution of modularity for regular communities in the 100 **runs of** C**OMBO**

Table 3: General statistics of communities identified in each view.

	Min Trad. Com.	Max Trad. Com.	Res. Com.	% Unstable
1K-ELITE	6	7	12	7.9
2K-ELITE	7	9	20	8.0
3K-ELITE	8	9	11	2.5
4K-ELITE	9	10	16	4.0
5K-ELITE	8	10	22	6.5
6K-ELITE	8	9	29	5.5
7K-ELITE	9	11	13	2.4
8K-ELITE	8	8	11	1.7
9K-ELITE	8	9	10	1.2
10K-ELITE	8	9	14	1.6

suggest they are very well meshed to the rest of the elite network.

We also compute modularity to evaluate the strength of the division of each view of the elite network into regular and resilient communities. Figure 6 shows the modularity of resilient communities and the distribution of modularity values for each run of COMBO in different views. With regards to modularity, a higher modularity shows a better grouping of the graph into tight modules. The figure shows that as the network is extended to cover more elites, COMBO is able to find tighter communities. The figure also shows that resilient communities are slightly less modular than the regular communities in certain views. For instance in 2K-ELITE, 5K-ELITE, and 6K-ELITE the modularity of resilient communities is approximately 0.04 lower compared to regular communities. This result, in addition to the findings in Figure 4, shows that the connectivity in this view exhibits less pronounced modular structure and has higher similarity to the connectivity in a random graph [6]. For the other view, however, the modularity of resilient and regular communities are very similar. We conclude that resilient communities each contain a group of nodes with a large number of social ties within the resilient community and a small number of friendships with users in other resilient communities.

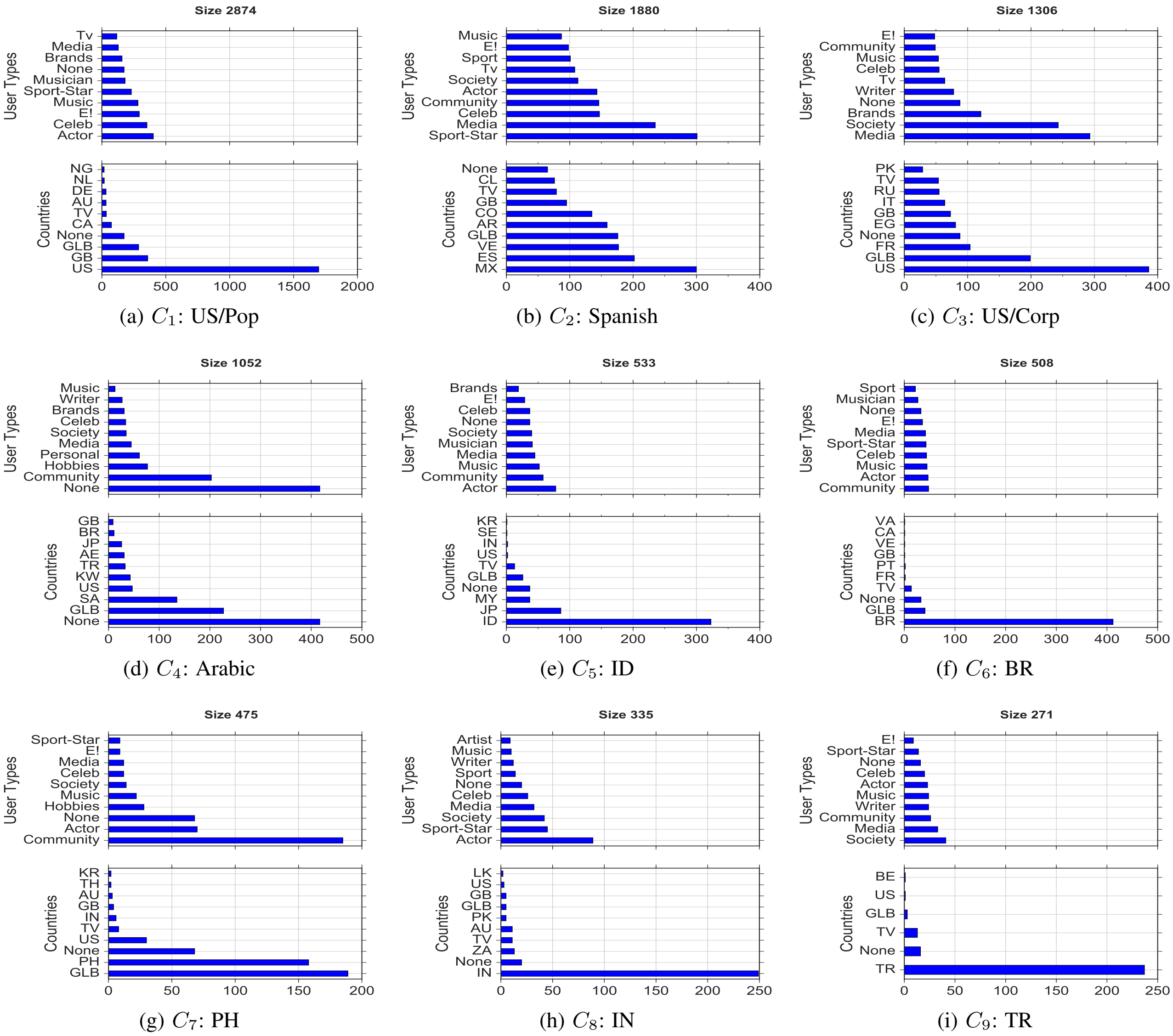
3.1 Social/Geo Cohesion of Elite Communities

Given the identified elite communities, the question is "*whether the elite communities represent meaningful units of the network?* We tackle this question by exploring whether users in each community exhibit social cohesion. We recall that socialbakers.com provides the 8 social categories (and 137 subcategories) as well as 196 unique countries as the location attribute for more than 90% of elite users. Using this information, we examine the histogram of these attributes across users in each elite community (*i.e.*, its social & geo footprints) to assess their level of social cohesion. Figure 7 shows the social footprint of three elite communities in 10K-ELITE view and the rest are available in the related technical report [16].

The examination of these footprints clearly shows that they all exhibit a significant level of social and/or geo (or language) cohesion. Since many elite accounts belong to easily recognizable individuals/entities, we manually inspect accounts in each community and leverage their social context to identify the "theme" associated with each community[4]. Table 2 summarizes the main features of the top 14 elite communities in 10K-ELITE, namely their assigned label, their size and their theme. While the level of cohesion might vary among communities, they all exhibit a very pronounced theme. We observe the role that imagined communities play in the Twitter elite Network. For example, there are some geopolitical and language-based communities (*e.g.*, Spanish, BR, ID) as well as other cultural-based communities US/TV, Adult, or US/Corp [9]. However, the concept of imagined communities does not explain the entire community structure of the elite network. For example, some of the categories are more difficult to classify (*e.g.*, GLB/-FUN) while unstable nodes are located between communities.

C_1 **- US Popstar** (2799)**:** This community is associated with celebrities, popstars and entertainment media. The vast majority of these elites are from the US with the remainder almost exclusively from English-speaking countries. US popstars, such as Katy Perry and Justin Bieber, and pop media programs, such as the Ellen Show and the X Factor, play a prominent role in this community. A noticeable teen or "tween" icon thread weaves through this community with Selena Gomez and Ariana Grande and with former Disney stars, such as Justin Timberlake, Christina Aguilera, Britney Spears, and Demi Lovato.

[4] The list of accounts associated with each elite community in samples views are available online at https://goo.gl/UGqcqf.



(a) C_1: US/Pop

(b) C_2: Spanish

(c) C_3: US/Corp

(d) C_4: Arabic

(e) C_5: ID

(f) C_6: BR

(g) C_7: PH

(h) C_8: IN

(i) C_9: TR

Figure 7: The distribution of category and country across accounts in the identified communities of the 10K-Elite.

C_2 **- Spanish Speaking** (1827)**:** A common theme across accounts in this community is its common language of Spanish. Geographically, 40% of these elites are from Mexico and 30% are from Spain. Yet, the geographic distribution draws from a wide swath of both Spanish-speaking elites with a small, but important group of non-Spanish speaking elites. Another theme which is less pronounced in this community is the focus on sports. This community consists of numerous globally popular soccer icons, such as Cristiano Ronaldo and Wayne Rooney, and sports organizations, such as FIFA and the Olympics, but also Spanish-speaking actors and popstars, such as the Columbian singer Shakira and Puerto Rican singer Ricky Martin.

C_3 **- US Corporate Celebrities & Media** (1234)**:** This community is associated with the US and Global media stars and corporate elites in the US and UK. This community consists of accounts associated with media groups, corporations and global entities. For example, this community consists of global news and media organizations, such as the BBC, the Guardian (the entire news family), Reuters, CNN, The Economist, all major TV channels in the US, and personalities such as Anderson Cooper and Piers Morgan. Global business leaders, corporations, and institutions are also central to this community, such as Bill Gates, Samsung Mobile, Unicef, Facebook, Google, and NASA. We refer to this community as "US/Corp". Interestingly, these elite are also interconnected with a small collection of Middle Eastern elites from several countries. For example, this community includes the Kuwaiti imam Mishary Bin Rashid Alafasy, Queen Rania from Jordan, in addition to the Lebanese popstar Nancy Ajram and the Emirati vocalist Ahlam. This central community is both the most obviously cosmopolitan consisting and the most corporate hinting at the global reach of Twitter, while also indicating that corporate world is deeply embedded within this facet of digital social life. This community includes US Popstar that are less geographically diverse than the previous two communities and many of the elites share similar categories, specifically singers and actors.

C_4 **- Arabic Speaking** (956)**:** This community mainly consists of Arab elites. Interestingly, these accounts mostly belong to media agencies and communities. We should note that the many of the elites in this community are not indexed in socialbakers.com, hence the most common country and user type in Figure 7(d) is *None*. However, we extract its social and language context by manually inspecting elites in this community. Mentionable famous Arab accounts in this community are Al-Arabiya and the Al-Jazeera news group.

C_5 **- Brazilian** (496)**:** Referred to as "BR", this community is almost entirely populated by Brazilian cultural elite individuals and organizations, such as the soccer stars Kaka and Neymar, and the television network, Rede Globo.

C_6 **- Filipino** (461)**:** Referred to as "PH", Most accounts in this community are celebrities from the Philippines. Although many accounts in this community are categorized as GLOBAL (see Figure 7(g)), close examination revealed that they are in fact Filipino.

C_7 **- Indonesian and Malaysian** (231)**:** Users in this community are mostly from Indonesia and Malaysia. Interestingly, the elites within this community represents a diverse selection of celebrities and communities. An example of a user in this community is Agnes Monica, the Indonesian popstar. We refer to this community as "ID".

C_8 **- Indian** (317)**:** Referred to as "IN", this community represents a range of Twitter accounts for cultural and political Indian elites. For example, the actor Amitabh Bachan, the cricket star Suresh Raina, and Narendra Modi, the Prime Minister of India, are in this community.

C_9 **- Turkish** (242)**:** This community consists of various categories of Turkish elites. Popular Turkish organizations, such as the soccer club Galatasaray, NTV television networks and online media celebrity Cem Ylmaz are in this community.

C_{10} **- K-Pop** (142)**:** This community mainly consists of Korean popstars. Among well known elites we can name the Korean actor Siwon Choi. Even non-Korean accounts within this community are focused around K-Pop (*e.g.*, @allkpop).

Other communities that each have less than 50 users, include *Thai* (28), *Adult* (20), *US TV stars* (19), and *Global fun* (13).

3.2 Communities Across Different Views

One important question is *"whether and how the social cohesion and theme of elite communities may vary across different views?"* To answer this question, we consider 10 different views of the elite network (1K-ELITE, 2K-ELITE, ..., 10K-ELITE), detect the resilient communities in each view, determine their social and location footprints. Furthermore, we keep track of the overlapping users between communities in consecutive views to establish their similarities. Figure 8[5] shows the relationships among communities in consecutive views as we extend the size of the elite network using a Sankey flow diagram [26]. The x axis shows the size of the elite network as it grows by 1K in each step and each group of vertically aligned bars represents communities in a particular view. The length of each bar indicates the size of the corresponding community and its label shows the name of the community using the following convention: view.size-theme. For example E9K-BR is a community in elite network of top 9K whose main theme associated with Brazil. The gray horizontal strips between communities in consecutive views show the number of overlapping users (and thus similarity of themes) between those communities. Figure 8 illustrates the following key points: First, the collection of main themes across communities in various views are rather stable. Furthermore, our closer examination also show that the elite communities at all views exhibit a very pronounced social cohesion. Second, as new nodes are added to the network, many communities remain relatively stable (*e.g.*, E*K-*-BR, E*K-*-IN) while others merge or split across different views. The former group often has a consistent theme that may evolve over time (*e.g.*, "E6K-US/Media" evolves to "E7K-US/Corp" or "MX-Celeb" changes to "Spanish") but in the latter group the theme of communities often narrows (or broadens) as they split (or merge) (*e.g.*, "E9K-ID" splits into a larger "E10K-ID" and a smaller "E10K-KPop", and "E6K-Spanish" and "E6K-ES/GB/Sport" merge to form "E7K-Spanish"). Third, the size of the elite communities increases as the elite network grows and their mapping across consecutive views becomes more clear (*i.e.*, the gray strips become wider and have less splitting between the last three views). *The relative stability of themes of elite communities across different views clearly indicate that these themes are not a side-effect of a particular network size and rather represent an inherent social footprint of these communities. This in turn confirms that elite communities with their specific themes are "socially meaningful" components and their connectivity present a coarse view of the elite network.*. Later in Section 6, we show how these communities can offer a coarse view for the entire Twitter structure. Given the relative stability of themes and communities across different views (in particular larger ones), we primarily focus on the largest elite network (10K-ELITE) for the rest of our analysis in this paper to keep the discussion more clear.

A plausible explanation from sociology for the relative stability of elite communities is homophily or the similarities between connected users[21]. While a more microscopic view of the network may exhibit more changes among connected users, a macroscopic community-to-community view isolates key macro characteristics that relate users in each community. The stability of this property in larger views is consistent with the fundamental role of homophily in communication networks[21].

4. COMMUNITY-LEVEL CONNECTIVITY

Presence of elite communities as meaningful components of the elite network allows us to examine its connectivity structure at a community level. This coarser view not only is more comprehensible but also shows the relationship among these communities. In this section, we explore the following two notions of pairwise connectivity for the 10K-ELITE network: *(i)* direct friend-follower relationships, and *(ii)* indirect reachability.

4.1 Direct Friend-Follower Relationships

A friend-follower relationship (*i.e.*, an edge) from user u to user v indicates that v is interested in following (and receiving tweet from) u. Similarly, the collection of such relationships from elites in community C_i to their followers in community C_j illustrates the collective attention that C_i receives from C_j. Therefore, a direct connectivity structure among all elite communities reveals larger patterns of interest across these units. We emphasize that all communities are interconnected. Our goal is to examine whether their connectivity exhibit any bias. The heatmap in Figure 9(a) illustrates the relative bias in directed connectivity between elite communities. More specifically, the color of cell (i,j) shows whether the number of directed edges from community i to community j is larger or smaller than the degree-preserving randomized version of the elite network[6]. Compared to the randomized structure, having more edges (shown in red) indicates a positive bias and having less edges (shown in blue) implies a negative bias. All communities are ordered based on their size from bottom-up on the y axis and from right to left on the x axis.

[5] The interactive version of this figure is available online at https://goo.gl/M66hyU

[6] In a randomized version of the network, we randomly connect elite nodes while maintaining their in- and out-degrees.

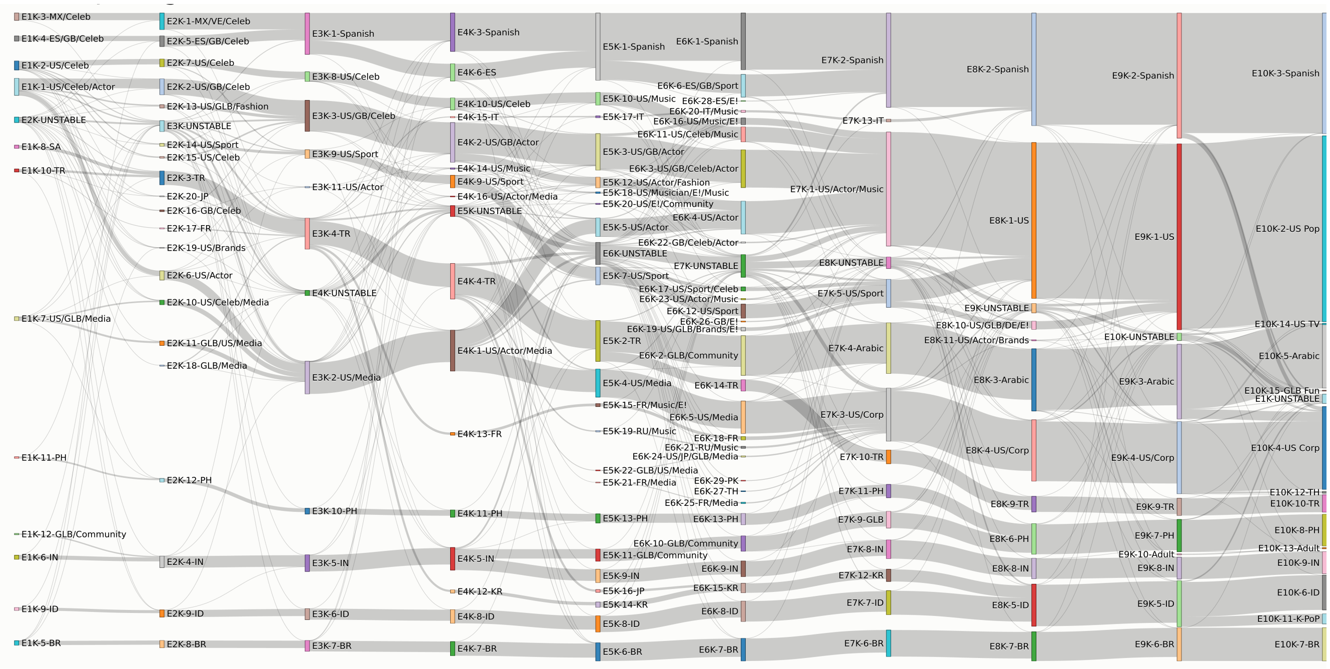


Figure 8: The evolution of elite communities and their themes across different views of the elite network from 1K-ELITE through 10K-ELITE

Figure 9(a) illustrates a few interesting points about the bias in connectivity between elite communities as follows: First, we observe that most cells on top and left side of the heatmap are white which indicates the lack of bias in their connectivity. Second, not surprisingly, there is a strong bias in intra-connectivity for larger communities (diagonal cells on the bottom right corner). Third, there is a strong negative bias in the connectivity between the four largest communities (bottom-right corner). US-Pop and Arabic communities particularly show a negative bias in their connectivity to other eight largest communities as well. Fourth, we clearly observe a reciprocal but mild positive bias on some off-diagonal cells namely between US-TV and UC-Corp, US-TV and US-Pop in both directions. Furthermore a few communities (US-Corp, Spanish, IN, PH) show a mild positive bias in their connectivity to unstable nodes.

4.2 Indirect Pairwise Reachability

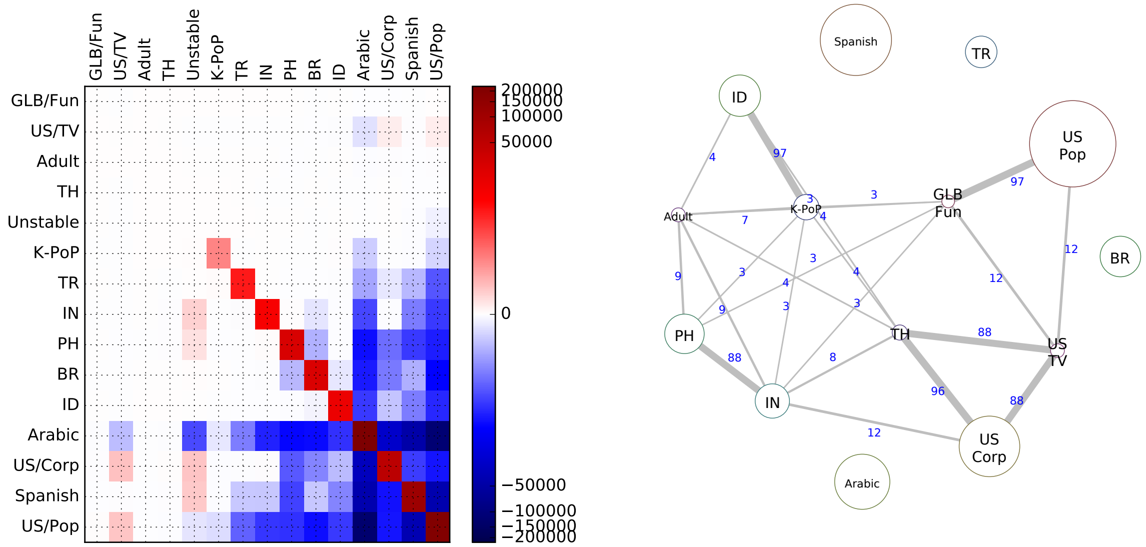
The "pairwise reachability" (*i.e.*, tight coupling) between two elite communities is an important aspect of connectivity that is not always correlated with the number of direct edges between them. In this section, we examine the notion of *pairwise reachability* between elite communities. To assess this rather subtle measure of connectivity, we examine the outcome of the individual runs of (Combo) community detection on the elite network. We recall that a detected community C_x in each run of combo may include two (or more) resilient communities RC_i and RC_j. Such a "co-appearance" of RC_i and RC_j is an indication of their relative reachability (or coupling). Therefore, the frequency of co-appearance for two resilient communities RC_i and RC_j in identified communities by Combo (across 100 runs in Section 3) is considered as a good measure to assess their pairwise reachability.

Figure 9(b) summarizes the pairwise reachability between all elite communities in 10K-ELITE where each circle represents a community. The thickness of each undirected edge between a pair of nodes shows their pairwise reachability. We also label each edge with the corresponding frequency of co-appearance for nodes at both ends. In essence, Figure 9(b) basically shows the likelihood of bundling between resilient communities in the outcome of each run of Combo.

This figure illustrates that most pairwise co-appearance frequencies are less than 13%. However, there are four distinct groups of elite communities that co-appear together much more frequently (>88% of time) as follows: *(i)* US-Corp, US-TV and Thai, *(ii)* ID and K-PoP, *(iii)* IN and PH, and *(iv)* US-Pop and Glb-Fun. Note that such a tight coupling between these communities were not apparent based on their direct connectivity in Figure 9(a) such as US-Corp and TH, ID and K-PoP and US-Pop and Glb-Fun. We can also observe that a few elite communities (*e.g.*, TR, BR, Arabic, Spanish) never co-appear with others which reconfirm their clear separation from other elite communities. The low frequency of co-appearance between two elite communities suggests that we can consider them as rather "unrelated/disconnected" components of the elite network. Therefore, we can conclude that the 10K-ELITE view of elite network consists of 10 separate components, the above four groups and six individual elite communities.

4.3 Role of Unstable Nodes

As we described in Section 3, unstable nodes do not consistently get grouped with any specific elite community since they have connections to users in many different clusters. This raises the following question: *whether unstable nodes serve as a bridge (or hub) between a pair/group of elite communities that result in their higher level of pairwise reachability?* To investigate this question, we consider all 155 (1.5%) unstable nodes in 10K-ELITE and determine their frequency of co-appearance with each elite community. Figure 10 presents the frequency of co-appearance of these unstable node as a heatmap where the color of the cell (i, j) indicates the frequency of co-appearance for unstable node j with elite commu-



(a) The level of bias in directed connectivity between elite communities in 10K-ELITE compared to its randomized version

(b) The frequency of co-appearance (pairwise reachability) of elite communities in 10K-ELITE

Figure 9: Characteristics of the community-level connectivity in the elite network

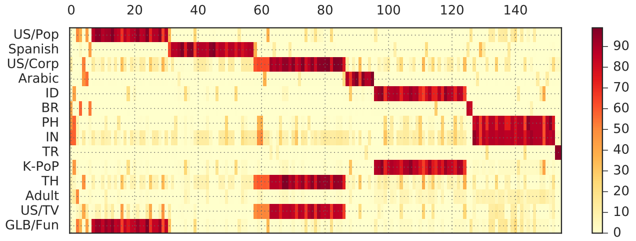


Figure 10: Pattern of co-appearance for 150 unstable nodes (on the x axis) with 14 elite communities (on y axis) in 10K-ELITE

nity i[7]. The higher frequency of co-appearance clearly indicates a tighter connectivity between the node and an elite community. Figure 10 reveals that four groups of these unstable nodes primarily co-appear with the following set of elite communities (from left to right): *(i)* US-Pop and Glb-Fun, *(ii)* US-Corp, TH, and US-TV, *(iii)* ID and K-PoP, *(iv)* PH and IN. These are exactly the same groups that showed significant pairwise reachability in Figure 9(b). This suggests that *these unstable nodes act as hubs and facilitate tighter coupling/reachability between the corresponding elite communities*. Figure 10 also shows two large groups of unstable nodes that are "hanging" from (*i.e.*, primarily co-appear only with) Spanish, and Arabic elite communities and two smaller ones from BR and TR communities.

5. INFLUENCE AMONG ELITES

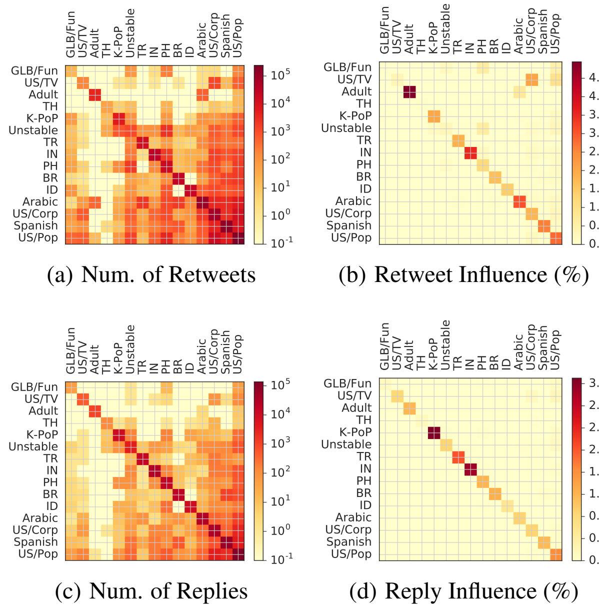
In this section, we investigate how elite communities influence each other. Prior studies on user influence have examined influence of user u on all other users in a social network using metrics such as the total number of retweets, mentions, or replies by other users on posts originated by u. While these measures of user engagements are user degree are generally correlated [10], the ranking of influential users based on user engagement and user connectivity measures (*e.g.*, PageRank) are not strongly correlated [12, 7]. There are four important differences between our analysis of cross-community influence and prior studies [7, 12, 2, 8] as follows: First, we only focus on influence between elite users (rather than all users) in a network. Second, we consider a modified version of an engagement-based metric based on *retweet* and *reply* to quantify pairwise influence between elite users. Third, we characterize cross influence at the granularity of elite communities rather than individual users. Fourth, we examine the relationship between community level influence and community level structure in the elite network.

Most prior engagement-based influence measures for user u use the total number of retweets or replies by all other users to u's post (*e.g.*, [7]). We capture the overall influence of elite user u (in terms of retweet or reply) on all other elites with the following two metrics: *(i) Number of influenced elites*: the number of unique users who have retweeted (or replied to) at least one of u's original tweets. *(ii) Aggregate influence*: This is the summation of the fractions of any other elite's captured tweets that are retweet of (or reply to) tweets originally generated by u. More specifically, the aggregate user influence of user u is defined as follows:

$$AggUserInfl(u) = \sum_{v \in Elite} \frac{RT_{u \rightarrow v}}{N_v} \tag{1}$$

where $RT_{u \rightarrow v}$ denotes the number of times that user v retweeted (or replied to) user u and N_v is the total number of v's tweets. We can also define the retweet (or reply) influence of community C_i on community C_j as a summation of all pairwise influences of users in C_i on users in C_j as follows:

[7] We use a simple reordering algorithm along the x-axis to group unstable nodes that have a similar co-appearance patterns. Note that the sum of the values in each column is not 100% since a co-appearance of an unstable node with multiple resilient communities is counted separately.



(a) Num. of Retweets (b) Retweet Influence (%)

(c) Num. of Replies (d) Reply Influence (%)

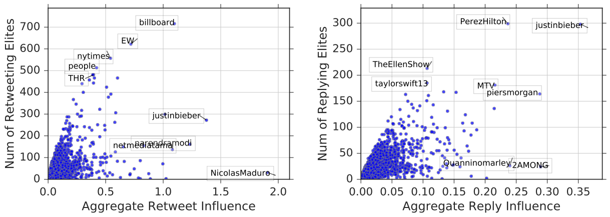
Figure 11: The absolute number of retweeting and replaying as well as retweet and reply influence across communities. The value in cell i, j presents the number or percentage of tweets (or replies) by users in community i to the posts that were originated by users in community j

$$AggCommInfl(C_i, C_j) = \sum_{v \in C_i} \sum_{w \in C_j} \frac{RT_{v \to w}}{N_W} \quad (2)$$

To conduct these analysis, we collect all available posts of all accounts in 10K-ELITE. Our datasets contains more than 31M tweets where 6.5M of them are retweets and 5M are replies.

The heatmaps in Figure 11 present two views of retweet and reply influence between elite communities. In Figure 11(a) (Figure 11(c)) the color of cell (i,j) indicates the absolute number of times that a user in community i has retweeted (replied to) tweets originated by users in elite community j. We observe that users in various communities retweet and reply to posts by users in other communities in particular those from US-Pop, Spanish and US-Corp. Figure 11(b) (Figure 11(d)) presents the normalized view of influence where the color of cell (i,j) indicates the percentage of tweets by user in community i that is a retweet of (reply to) tweets originated by users in elite community j. This normalized view properly represent the influence between elite communities and has non-zero values on the diagonal cells. This clearly demonstrate that *both the retweet and reply influence between elite users are primarily contained within their own community.* The only noticable exception to this clear pattern is the retweet influence of US-Corp on US-TV. Furthermore, the level of influence within elite communities vary. Elite users in Adult, Arabic and IN have the most retweet influence while those in K-PoP, PH, and IN show the most reply influence on their community members.

To gain more insight, we also characterize the patterns of pairwise user-level influence among elites. Figure 12 depicts both dimensions of influence for individual elite users in a scattered plot where each point presents a user, its x value indicates user's aggregate retweet (or reply) influence and its y value shows the number of unique elites influenced by the user. In Figure 12(a), we observe that elite users that are influential with respect to retweeting (*e.g.*, @nicolasmaduro, President of Venezuela) often influence a small number of elite users. Other accounts with a lower retweet influence (*e.g.*, @justinbieber, @billboard) tend to influence a much larger number of elites.



(a) Retweet (b) Reply

Figure 12: Visualizing the two dimensions of influence for individual elite users based on both retweet and reply measure.

Table 4: Top 10 most influential elites in the 10K-ELITE based on different metrics: PageRank, the number of retweeting or replying elites

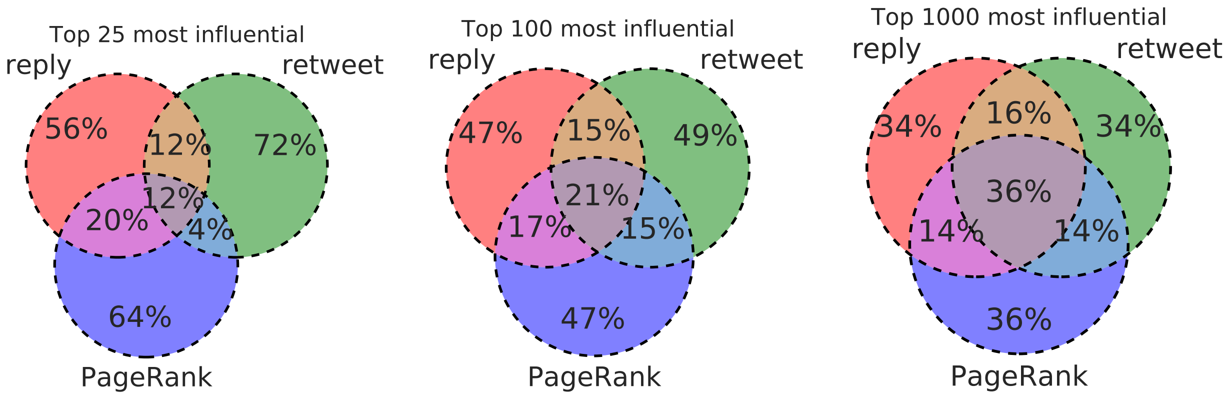
Rank	PageRank	Reply	Retweet
1	instagram	PerezHilton	billboard
2	mikeyk	justinbieber	EW
3	twitter	TheEllenShow	nytimes
4	BarackObama	taylorswift13	people
5	Pontifex	MTV	Variety
6	Pontifex_es	edsheeran	THR
7	Pontifex_pt	realDonaldTrump	TIME
8	Pontifex_it	piersmorgan	AppleMusic
9	nytimes	jimmyfallon	mashable
10	jimmyfallon	KimKardashian	RollingStone

Examination of the reply influence in Figure 12(b), shows that users exhibits different combination of both dimensions and may have a large value in both dimensions a rather different characteristics as we can identify users that exhibit (*e.g.*, @PerezHilton and @justinbieber).Furthermore, both the aggregate influence and the number of influenced users are smaller for retweet than reply measures. *These analysis illustrates that both dimensions of retweet or reply influence are equaly important to gain a complete picture of the pairwise influence between elite users*

Finally, examining the usernames of reply influential accounts show that they often belong to celebrities in the entertainment industry and gossip media, *e.g.*, @PerezHilton (the gossip blogger and columnist) and @justinbieber (the popstar singer).

Table 4 summarizes the top-10 most influential elites based on their PageRank and one measure of retweet or reply influence, namely the number of influenced elites. We observe that except for @jimmyfallon, who appears in two top-10 rankings, there is no other overlap between them.

The observed minimal overlap among the top-10 most influential users based on different measure raises the following question: *"How does the overlap among the top-N most influential users based on different metrics change with N?"* Exploring this question reveals the level of separation between the influential users according to each measure. The three Venn diagrams in Figure 13 present pairwise and three-way overlap among top-N influential users according to the three metrics for N equal to 25, 100 and 1K. We observe that the 3-way overlap among different groups of influential



(a) Ranked as top 25 (b) Ranked as top 100 (c) Ranked as top 1K

Figure 13: Overlap among different influence measures

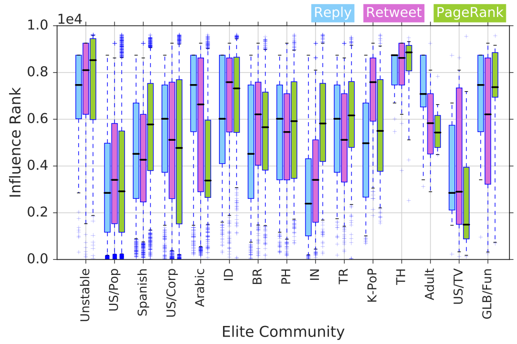


Figure 14: Distribution of the rank of users in each elite communities in 10K-ELITE based on two measure of influence and PageRank

users grows with N from 12% to 21% and 36%. Interestingly, even for the top-1K, between 34-36% of users are considered influential based on just a single metric, and the plots do not reveal any similarity between ranking observed by any two metrics. Hence, each of these metrics captures a different aspect of importance/influence, and the topological centrality of a user's position in the social graph does not lead to his success in attracting large reactions from other elites. This finding is generally aligned with the lack of correlation in various measurements of influence in prior studies [7, 12].

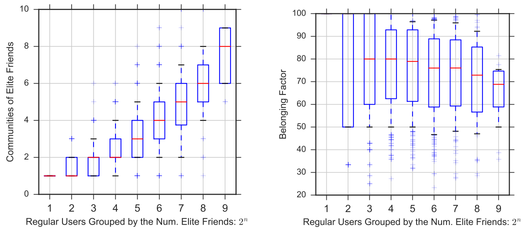
To get a broader view of influence for individual elite communities, we examine the influence of their nodes based on different metrics. Figure 14 presents the summary distribution of rank among all elites in 10K-ELITE based on our two measures of influence (*i.e.*, retweet and reply) for users in each elite community (including unstable nodes). Furthermore, we also include the summary distribution of user ranks based on its PageRank [5] in the elite network as an overall measure of centrality for each elite community. Note that each one of these summary distribution of ranks for users in an elite community demonstrate a different aspect of their influence. This figure shows that the relative ranking of users based on different influence measures have rather comparable ranges for most elite communities. The exceptions are Arabic, IN, K-PoP and Adult communities that exhibit very different ranking for various influence measures. For example, users in IN community have a high reply influence, moderate retweet influence but low centrality ranking. Users in the Arabic community show an opposite pattern with a higher centrality ranking, much lower retweet ranking and even lower reply ranking. These patterns basically reflect the nature of overall influence of an elite community on the rest of elite network. Finally, we observe that US-Pop and US-TV have the highest overall influence whereas TH, Glb-Fun, Adult and ID have lowest engagement influence on other elites.

6. FROM COMMUNITIES TO PARTITIONS

In our analysis in Section 3, we showed that the themes of the 14 identified elite communities are pretty visible in the elite network at different sizes. The persistent visibility of elite communities with specific themes at different size of the elite network raises the following question *whether the remaining regular (*i.e.*, non-elite) users can be divided into mutually exclusive groups, each one representing the extension of an elite community?*. In particular, we focus on 80+% of regular users who follow at least one elite user in 10K-ELITE (*i.e.*, have an elite friend) and examine whether these nodes can be partitioned based on their association with a particular elite community.

In order to establish an association between regular users and elite communities, we select 10K random regular users [8] since it is not feasible to consider all regular users. To relate a regular user to a specific elite community based on two observations as follows: First, we group regular users based on their number of elite friends into exponential buckets (*i.e.*, bucket x contains users whose number of friends is between 2^x, 2^{x+1}-1). Figure 15(a) shows the summary distribution of the number of elite communities where the elite friends of regular users in each group are located. This figure clearly illustrates the number of elite communities that a regular user follows logarithmically increases with the number of its elite friends. Second, the fraction of elite friends of a regular user u that are located in elite community c can be viewed as u's *belonging factor* with community c. Then, we map user u to the elite community that has the largest belonging factor (*i.e.*, contains most of its elite friends. Using the exponential grouping similar to Figure 15(a) , we show the summary distribution of largest belonging factor for regular users in each group in Figure 15(b). We observe that more than 70% of elite friends of regular users are typically located in the same elite community. This rather skewed association of individual regular user to elite communities suggest that each

[8] These random users are identified using random walks from a randomly selected nodes.



(a) Distribution of number of elite communities that a regular user follows

(b) Distribution of the belonging factor

Figure 15: Using elite communities as landmarks to cluster regular users.

user can be rather reliably mapped to the elite community with the largest belonging factor.

We argue that a collection of regular users that are mapped to a single elite community, can be viewed as a "shadow partition" of that community. To support this claim, we consider 100K randomly selected friend-follower relationships between regular users and then map the regular users at both ends to their corresponding elite communities. We observe that 35.2% of these relationships are between users in different shadow partitions. This is very similar to the fraction of relationships between elite users that are located in different elite communities. *The similarity of the fraction of relationships between elite users that are in different communities with the fraction of relations between regular users that are in different shadow partitions indicates that each shadow partition is an extension of its corresponding elite community,* i.e.*, elite communities can be used to partition regular users and thus they present a coarse view of the entire Twitter structure.*

7. CONCLUSION & OUTLOOK

In this paper, we present a socially-informed approach to characterize the structure of connectivity among Twitter users. We rely on the sub-graph of most-followed users (or elites), called elite network, since it serves as the backbone of the structure. We present a new technique for capturing and validating the Twitter elite network. We use this technique to capture Twitter elite network and annotate each node with its social attributes for our analysis. We show modularity-based communities of elites exhibits social cohesion with a clear theme across different sizes of elite network. This is a strong indication that elite communities are socially meaningful. We then characterize both the connectivity and influence among elite communities. For example, we show that users in different elite communities have a minimal influence on each other. Finally, we illustrate the tendency of regular users to follow elites in a single elite community and suggest to group regular users into "shadow partitions" based on their interest to an elite community. Our examination shows that the fraction of relationships between elite that span across communities is very same as the fraction of relationships between regular users that span across different shadow partitions. This indicates that each shadow partition can be viewed as an extension of its favorite elite community. Our findings collectively demonstrate that elite communities represent a coarse view of the elite network as well as the entire Twitter structure.

Some of our plans to extend this work are as follows: First, we will examine whether there are other evidences (*e.g.*, any context from regular users) to confirm the association between each elite community and its corresponding shadow partition. Furthermore, it is worth exploring whether there are size and structural similarities between elites communities and their shadow partitions. Finally, we plan to extend the notion of shadow partitions by leveraging individual elite communities as landmarks and cluster regular users based on their level of connectivity to all elite communities.

8. REFERENCES

[1] K. Avrachenkov, N. Litvak, L. O. Prokhorenkova, and E. Suyargulova. Quick detection of high-degree entities in large directed networks. In *Proc. of ICDM*. IEEE, 2014.

[2] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proc. of WSDM*. ACM, 2011.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), 2008.

[4] B. Bollobás. *Modern graph theory*, volume 184. Springer Science & Business Media, 2013.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of WWW*, 1998.

[6] R. Campigotto, J.-L. Guillaume, and M. Seifi. The power of consensus: random graphs have no communities. In *Proc. of ASONAM*. ACM, 2013.

[7] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 2010.

[8] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proc. of WWW*. ACM, 2009.

[9] R. Dunbar, V. Arnaboldi, M. Conti, and A. Passarella. The structure of online social networks mirrors those in the offline world. *Social Networks*, 43:39–47, 2015.

[10] D. Easley and J. Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[11] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3), 2010.

[12] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? In *Proc. of WWW*. ACM, 2010.

[13] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community Structure in Large Networks: Natural Cluster Sizes and The Absence of Large Well-Defined Clusters. *Internet Mathematics*, 6(1), 2009.

[14] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis. Tastes, ties, and time: A new social network dataset using facebook. com. *Social networks*, 30(4):330–342, 2008.

[15] R. Motamedi, R. Rejaie, D. Lowd, and W. Willinger. WalkAbout: Exploring the Regional Connectivity of Large Graphs and Its Application to OSNs. Technical report available at: http://onrg.cs.uoregon.edu/pub/tr13-06.pdf, University of Oregon, 2014.

[16] R. Motamedi, S. Rezayi, R. Rejaie, R. Light, and W. Willinger. Characterizing twitter elite communities: Measurement, characterization, and implications. Technical report available at: http://onrg.cs.uoregon.edu/pub/tr16-15.pdf, University of Oregon, 2016.

[17] D. Murthy. *Twitter: Social communication in the Twitter age*. John Wiley & Sons, 2013.

[18] M. E. Newman. Modularity and community structure in networks. *Proc. of NAS*, 103(23), 2006.

[19] D. Quercia, L. Capra, and J. Crowcroft. The social world of twitter: Topics, geography, and emotions. *ICWSM*, 12:298–305, 2012.

[20] A. H. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach. Respondent-driven sampling for characterizing unstructured overlays. In *Proc. of the INFOCOM*. IEEE, 2009.

[21] E. M. Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

[22] M. Rosvall and C. Bergstrom. Maps of information flow reveal community structure in complex networks. In *Proc. of the NAS*. Citeseer, National Academy of Sciences, 2007.

[23] S. Sobolevsky, R. Campari, A. Belyi, and C. Ratti. General optimization technique for high-quality community detection in complex networks. *Physical Review E*, 90(1), 2014.

[24] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On unbiased sampling for unstructured peer-to-peer networks. *IEEE/ACM Transactions on Networking (TON)*, 17(2), 2009.

[25] Wikipedia. Imagined community. https://en.wikipedia.org/w/index.php?title=Imagined_community&oldid=741155614, 2016. Accessed: 2016-10-24.

[26] Wikipedia. Sankey diagram. https://en.wikipedia.org/w/index.php?title=Sankey_diagram&oldid=740785912, 2016. Accessed: 2016-10-24.

[27] J. Yang and J. Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proc. of WSDM*. ACM, 2013.