2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

On Characterizing the Twitter Elite Network

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

Walter Willinger
NIKSUN, Inc.
wwillinger@niksun.com

***Abstract*—The most-followed Twitter users and their pairwise relationships form a sub-graph of all Twitter users that we call the *Twitter elite network*. The connectivity patterns and influence (in terms of reply and retweet activity) among these elite users illustrate how the "important" users connect and interact with one another on Twitter. Such an elite-focused view also provides valuable information about the structure of the Twitter network as a whole.**

This paper presents the first detailed characterization of the top-10K Twitter elite network. We describe a new technique to efficiently and accurately capture the Twitter elite network along with social attributes of individual elite accounts. We show that a sufficiently large elite network is typically composed of 15-20 resilient and socially cohesive communities representing "socially meaningful" components of the elite network. We then characterize the community-level structure of the elite network in terms of bias in directed pairwise connectivity and relative reachability. We demonstrate that both the retweet and reply activity between elite users are effectively contained within individual elite communities. Finally, we illustrate that a majority of the elite friends of regular Twitter users tend to belong to a single elite community. This finding offers a promising criterion to group regular users into "shadow partitions" based on their association with elite communities. We show that the level of overall inter-connectivity between shadow partitions mirrors the inter-connectivity of the elite communities. This suggests that these shadow partitions can be viewed as extensions of their corresponding elite communities.

I. Introduction

The increasing popularity of online social networks (OSN) such as Twitter has fueled the growing interest in characterizing their connectivity structure, the exchange of information among their users, and how their users influence one another. These studies usually focus on the entire connectivity structure [4], [13]. However, a majority of users in an OSN typically have a low level of connectivity and/or activity (*e.g.*, having only a handful of followers or tweeting just a few times a month). Therefore, any analysis of the entire network tends to be dominated by these regular users.

We argue that high-degree and/or highly-active or influential (or "elite") users [1] in an OSN play a significantly more important role in terms of connectivity, information propagation and influence than regular ("non-elite") users. Therefore, characterizing the connectivity structure of the *elite network*, the core subgraph of an OSN that contains all the "elite" nodes and their pairwise relationships offers a number of promising opportunities. First, characterizing the elite network has the potential of revealing the relationship and influence patterns between the elite users in a particular OSN. Such findings can offer valuable insights into how the OSN is used by these important users. Second, elite users are usually well-known individuals/entities (*e.g.*, celebrities, news agencies, politicians) with specific social, cultural, or geographic attributes that can be leveraged to examine relationship patterns among elite users in a *socially-informed* manner. Lastly, since elite users collectively have direct connectivity to a significant number of regular users in an OSN, identifying certain relationship patterns among elite users has the potential of providing valuable insights into the overall structure of the entire OSN network. The key challenge in characterizing the elite network of an OSN is to accurately capture the elite network at the "right" level of granularity (*e.g.*, type of users and type of activities).

Motivated by these observations, this paper presents the first detailed characterization of the *Twitter elite network*. To this end, we first describe our proposed methodology for accurately and efficiently capturing the Twitter elite network that consists of the top 10K most-followed Twitter users (10K-ELITE) along with their social attributes and their pairwise follower-friend relationships. After identifying "elite" communities in the resulting elite network, we examine social attributes of the elite users in each of these communities to determine whether they exhibit any specific theme. We illustrate how elite communities grow, split, and merge as we grow the elite network from 1K to 10K nodes. The identified elite communities enable us to examine the structure of the Twitter elite network at the community level. More specifically, we study the pairwise connectivity between elite communities using two metrics that measure the level of biased in directed edges between them and their pairwise reachability, respectively. Based on these metrics, we *(i)* assess the influence of both retweets and replies at the user- and community-level in the elite network, and *(ii)* investigate the alignment between these measures of influence and the "importance" of users in the elite network measured by PageRank [6]. Finally, we study the connectivity between regular and elite users to determine whether elite communities can be leveraged to cluster the regular Twitter users in a meaningful and cohesive manner.

Our characterization of the Twitter elite network results in a number of specific findings. First, irrespective of the used cut-off (*e.g.*, top 10K most-followed users), all nodes in a Twitter elite network form a weakly connected component

[1] This material is based upon work supported by the National Science Foundation under grant IIS-0917381 and CNS-1320977.

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 © 2018 IEEE

that has a star-shaped structure. The largest strongly connected component (LSCC) that contains a vast majority of all the nodes and edges is at the center of this structure and is typically surrounded by singleton nodes that have directed edges to (*i.e.*, being followed by users in) the LSCC. Second, the Twitter 10K-ELITE network is composed of some 14 *elite communities* of different sizes that exhibit "social cohesion" around a common theme related to, for example, a country, a language, a cultural background, or a business interest. Furthermore, the number of elite communities and their associated themes remain rather stable once the elite network reaches a certain size (i.e., some 6K nodes). This observation suggests that elite communities present robust and socially meaningful entities of the Twitter elite network. Third, examining the obtained elite communities in more detail, we observe a symmetric negative bias in the directed connectivity between the four largest elite communities and notice a significantly higher reachability between a small subsets of communities. This higher reachability can be explained by a subset of elite users that are not part of any elite community but act as "bridges" between different elite communities. Fourth, we find rather surprisingly that for most elite communities, the normalized retweet and reply influence is primarily due to the elite users in the individual elite communities. Finally, we observe that a majority of the elite friends of regular Twitter users tend to belong to a single elite community. This finding suggests a promising criterion for grouping regular users into "shadow partitions" based on their association with users in elite communities. In particular, we show that the overall inter-connectivity between these shadow partitions closely mirrors the inter-connectivity between the corresponding elite communities which in turn suggests that the identified shadow partitions can be viewed as extensions of their corresponding elite communities.

The rest of the paper is organized as follows: In Section II, we present our technique for capturing the Twitter elite network. Our approach for detecting elite communities and identifying their basic characteristics is descried in Section III. Aspects of the inter-connectivity and cross-influence among elite communities are discussed in Section IV and V, respectively. In Section VI, we explain how the association of regular Twitter users with individual elite communities can be leveraged to group those regular users into shadow partitions. We conclude in Section VII by summarizing our contributions and outlining our plans for future work on this topic.

II. CAPTURING ELITE NETWORK

Our goal is to efficiently capture the Twitter elite network - that is a subgraph of Twitter that contains the top-N most-followed accounts (*i.e.*, nodes) 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 in order to use this annotated graph as input 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 using online resources and complementing that with customized random walks to discover more accounts, *(ii)* Identifying the pairwise connections between these accounts, *(iii)* Detecting any missing elite accounts and collecting their information, *(iv)* Collecting all profile information and available tweets of the elite accounts. All the data collection (*i.e.*, connectivity information and tweet activity) was performed in September of 2016. Next, we describe each of these four steps in more detail.

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 the most followed accounts in a variety of OSNs in different social categories (*e.g.*, celebrities, actors, sport, community, ...). Each list on `socialbakers.com` provides up to 1 000 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.

To independently identify Twitter accounts with many followers, we also conduct 2K "customized" random walks that start from randomly selected Twitter accounts. Our random walks only select a random user from the friend list of the current user as their next step. The likelihood that these walkers visit a user is proportional to its number of followers. Therefore, these random walks offer an efficient technique to identify the most-followed and visible users[2] [18]. We merge all the discovered accounts from our random walks with the accounts captured from `socialbakers.com` and mainly focus on the top 10K accounts with the most followers to form a *master list*.

Step 2: It is prohibitively expensive to find all the pairwise connections between the identified accounts by collecting and examining hundreds of millions of their followers that are mostly regular users. Our observation is that the number of friends for elite accounts are almost always several orders of magnitude smaller than the number of followers. Therefore, our key idea is to collect the complete list of friends (instead of followers) for each selected elite 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 reduces the overhead of capturing all

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

[2] A user with many followers that is part of a partition or weakly connected region is not likely to be discovered by random walks. We argue that such an elite user is less important for our analysis.

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 directed subgraph that connects the most-followed Twitter accounts. Since it is possible that the identified top-10K accounts in step 1 do not accurately represent the actual top-10K accounts on Twitter, we perform one more check to verify whether the list of identified account is correct and complete. We observe that any missing elite account is very likely to be followed by many elites that we already identified as top 10K accounts [2]. Since we already collected the entire list of friends for top-10K accounts, we can 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 in our list, 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 lists. Using this technique, we detected 264 (2.6%) missing elite accounts that are between the rank of 500 and 10K. The small percentage of the discovered missing accounts in this step along with their relatively low ranking indicate that our master list is reasonably accurate and complete. *In summary, among the top* 10*K most-followed Twitter accounts,* $8,704$ *are exclusively reported in* `socialbakers.com`*,* 301 *are found exclusively using random walks,* 731 *are confirmed by both techniques, and* 264 *are among the discovered friends of most-followed accounts.*

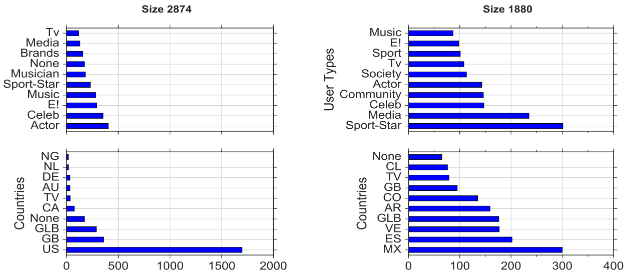
Step 4: We collect all the available tweets (up to the last 3 200) for each top 10K Twitter account. These tweets are used to investigate the influence between elites by analyzing retweets, and to gain some insight on how they use Twitter by analyzing tweets/retweets and constructing topic models.

Basic Characteristics of Elite Networks: While it is compelling to consider Twitter users with the highest number of followers as Twitter elites, one remaining question is *how many of the most-followed accounts should be considered as a part of the elite network?* We argue that the 10K-ELITE offers a sufficiently large view of the elite network in Twitter. For one, 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 significantly lower. Second, examining the friend lists of 10K random twitter users reveals that 80% of these random (and thus 80% of all) twitter accounts follow the top 10K elites. 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 a diminishing rate of return.

We examine *whether and how the size of the resulting elite network affects its structural properties* by considering 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. As the size of the elite network is extended (from 1K to 10K), it becomes denser (average degree increases from 49 to 152), but the fraction of reciprocated edges remains between 32-40%, which is higher than the reported 22% for the entire Twitter social graph [11]. 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) [9] in each view contains 91-95% of nodes and 94-97% of all edges in the elite network. Most of the the other strongly connected components (SCCs) consist of a single node while few of them have two or more nodes. In all views, all SCCs form a "star-like" structure where the LSCC is in the center and there are directed edges from other SCCs to nodes in the LSCC (*i.e.*, only elites in the LSCC follow and receive tweets from elites in other SCCs).

III. ELITE COMMUNITIES

Our next goal is to determine whether the elite network is composed of a collection of meaningful components that can be used to gain insight into the relationship among these elite users as well as the overall structure of Twitter. While the most natural components are groups of tightly connected nodes (or "communities"), applying out-of-box community detection algorithms is problematic. First, most commonly-used community detection techniques take undirected graphs as input while the elite network is a directed graph [10]. To address this issue, we first convert each view of the elite network into a weighted undirected graph by replacing *each* directed edge into a single undirected edge with the weight of 2 when reciprocal directed edges exist or the weight of 1 otherwise. This representation allows us to encode tighter bindings between users with reciprocal edges as compared to prior studies (*e.g.*, [12] simply converts a directed graph into an undirected one). Second, the outcome of some of the most commonly-used community detection techniques (*e.g.*, Louvain [4], BigCalmm [19], InfoMap [16], EDA [15]) is non-deterministic and varies across multiple runs. To address this issue, we use the COMBO community detection technique [17] that relies on multi-objective optimization and detects more stable communities across different runs as compared to, say, Louvain [4]. We also eliminate the residual instability by only considering a group of nodes as a community if they consistently grouped together across different runs. To achieve this objective, we run COMBO on each view of the elite network k times and collect the communities that individual nodes are mapped to in each run in vectors with k values,



(a) US/Pop (b) Spanish

Fig. 1. The social and geo footprints for three sample elite communities in 10K-ELITE.

the so-called the "community vectors". Then, we group all the nodes that are mapped to the same community in all runs (*i.e.*, have the same community vector) and refer to the group as a *resilient community*. This process of detecting resilient communities also results in group of nodes with unique community vectors. We group this set of nodes along with nodes in resilient communities that are smaller than 10 nodes in size and refer to them as set of *unstable nodes*.

Clearly, increasing k is more restrictive and may lead to smaller resilient communities since more runs can simply split a community into two (or more) smaller ones. 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. COMBO detects between 10-29 resilient communities across different views of the Twitter elite network; collectively, they cover 92-99% of all nodes in each view. Thus, less than 8% of the elite users are *unstable* nodes. We emphasize that the identified elite communities are different from typical communities that one obtains by running community detection on the entire Twitter graph that contain many regular (*i.e.*, non-elite) users.

Cohesion of Elite Communities: An important question is *whether the identified elite communities represent meaningful units of the Twitter network?* We answer this question by exploring whether users in each community exhibit social cohesion. We recall that socialbakers.com provides 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 the social and geographic attributes (*i.e.*, footprint) across users in each elite community to assess their level of social cohesion. Figure 1 shows the social footprints of two elite communities in the 10K-ELITE view. The footprints for other elite communities are shown in our related technical report [14].

Our careful examination of these footprints shows that all of them exhibit a significant level of social and/or geographic (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. Table I summarizes the main features of the top 14 elite communities in 10K-ELITE, namely their assigned label, their size and their theme along with their density and conductance[5].

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 I
LABEL AND KEY FEATURES OF 14 ELITE COMMUNITIES IN THE TWITTER ELITE NETWORK

While the level of cohesion varies among communities, all the communities exhibit a very pronounced theme; the themes can be broadly divided into four categories such as*(i)* Elites from a single country (ID, BR, PH, TR, TH), *(ii)* Elites from different countries with a common language (Spanish, Arabic), *(iii)* Elites with a similar cultural interest (US/Pop, K-PoP, Adult), and *(iv)* Entities from a similar business sector at a single or multiple countries (US/Corp, US/TV, Global/Fun).

Communities Across Different Views: We are also interested in *whether and how an elite community's social cohesion and/or theme varies 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, and determine their social and location footprints. Furthermore, we keep track of the overlapping users between communities in consecutive views to establish their similarities. Leveraging a Sankey flow diagram, Figure 2 shows the relationships among communities in consecutive views as we expand the size of the elite network. 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 elite 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*; *e.g.*, E9K-BR is a community in elite network of top 9K whose main theme is 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 2 illustrates that the collection of main themes among communities stabilizes across a few larger views of the elite network. We also observe that in general, elite communities exhibit strong social cohesion at all views. Figure 2 also shows that 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. While the former often have a consistent theme that may evolve over time (*e.g.*, "E6K-US/Media" evolves to "E7K-US/Corp" or "MX-Celeb" changes to "Spanish"), for the latter

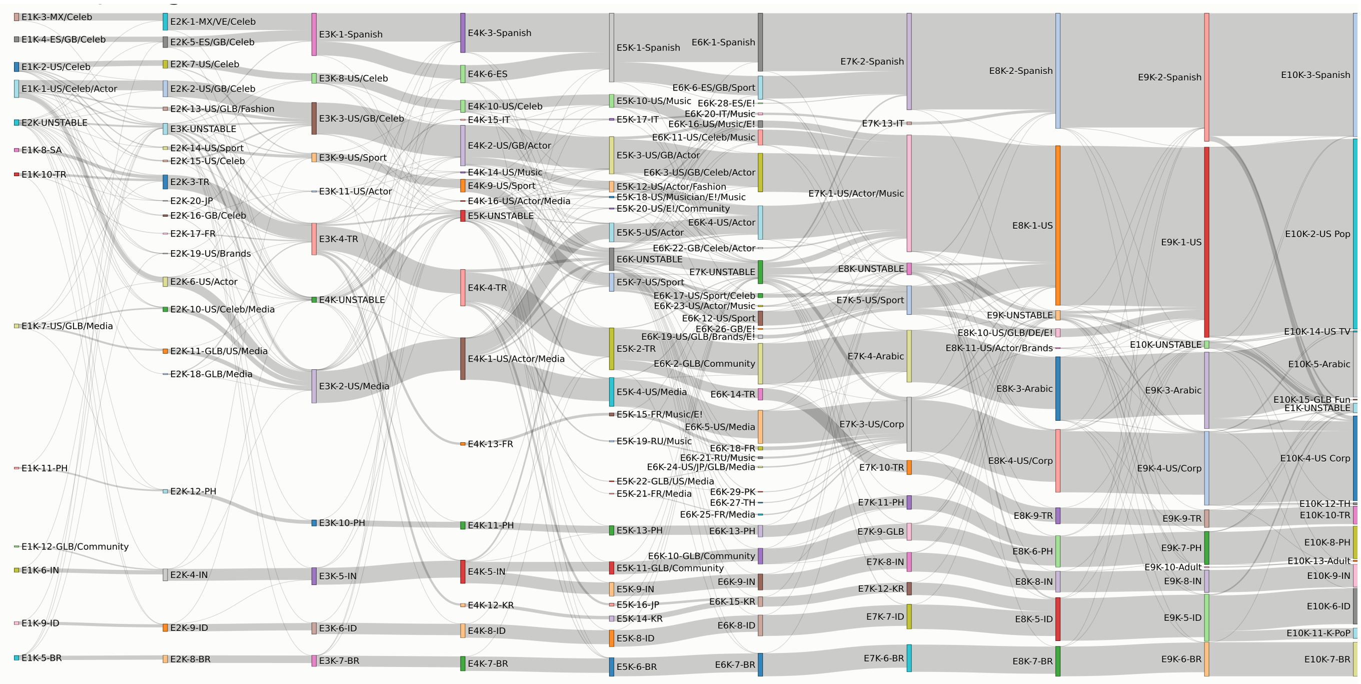


Fig. 2. The evolution of elite communities and their themes across different views of the elite network from 1K-ELITE through 10K-ELITE

the theme 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"). In addition, the sizes of the elite communities increase 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). To summarize, *the relative stability of themes of elite communities across different views clearly indicates that these themes are not a side-effect of a particular network size but instead represent a robust social footprint of these communities. This in turn confirms that elite communities with their specific themes represent "socially meaningful" components of the Twitter network.*

IV. COMMUNITY-LEVEL STRUCTURE

The presence of pronounced communities with social cohesion in Twitter's elite network allows us to examine its connectivity structure at the community level. This coarse-grained view of the Twitter network is not only more manageable in terms of scale but also more viable to study in terms of the relationships among these communities. To this end, we explore in this section the following two notions of pairwise connectivity for the 10K-ELITE network: *(i)* direct friend-follower relationships, and *(ii)* indirect reachability.

Bias in Directed Pairwise Connectivity: A friend-follower relationship (*i.e.*, an edge) from user u to user v indicates that v is interested in following (and receiving tweets from) u. Similarly, the collection of such relationships from elite users in community C_i to their followers in community C_j illustrates the collective attention that C_i receives from C_j. Therefore, a directed connectivity structure among all elite communities reveals larger patterns of interest across these units. We emphasize that there are edges between all pairs of elite communities. Our goal is to examine *whether the pairwise connectivity between different pairs of elite communities exhibits any bias*. The heatmap in Figure 3 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 C_i to community C_j is larger or smaller than the number of connections in a degree-preserving randomized version of the elite network[3]. 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.

Figure 3 shows that most cells in the top-left portion of the heatmap are white which indicates a lack of bias in their connectivity. At the same time, there is a strong bias in intra-connectivity only for the larger communities (diagonal cells in the bottom right corner) and a strong negative bias in the pairwise connectivity between the four largest communities (bottom-right corner). In particular, the US-Pop and Arabic communities exhibit a pronounced negative bias in their connectivity to all of the eight largest communities. Interestingly, all instances of a pairwise negative bias between communities are very symmetric. In addition, Figure 3 also shows a reciprocal but mild positive bias for some off-diagonal

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

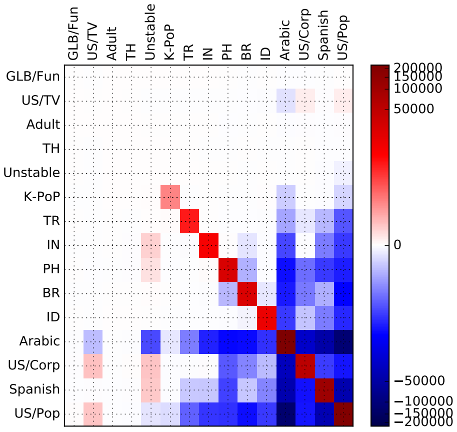


Fig. 3. Bias in directed connectivity between elite communities in 10K-ELITE

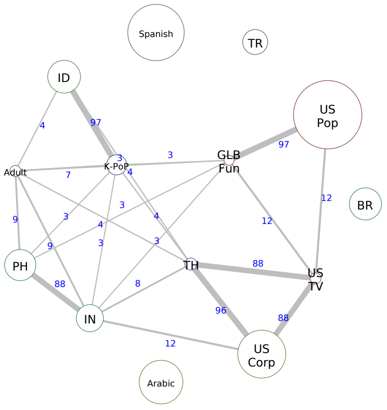


Fig. 4. The frequency of pairwise co-appearance of elite communities

cells; *e.g.*, between US-TV and US-Corp, US-TV and US-Pop in both directions. We also notice a few communities (US-Corp, Spanish, IN, PH) with a mild positive bias in their connectivity to unstable nodes.

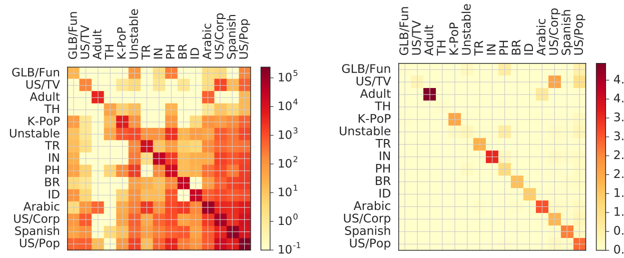
Indirect Pairwise Reachability: The "pairwise reachability" (*i.e.*, tight coupling) between two elite communities is an important aspect of connectivity that is not always correlated with the number of direct edges between them. To examine the notion of *pairwise reachability* between elite communities, we examine the outcome of the individual runs of the (Combo) community detection algorithm 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 communities identified by Combo (across 100 runs in Section III) can be considered to be an informative measure for assessing their pairwise reachability.

Figure 4 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 4 shows the likelihood of bundling between all pairs of resilient communities in the outcome of each run of Combo. Figure 4 reveals a few interesting points. First, a few elite communities (namely TR, BR, Arabic, Spanish) never co-appear with others which reconfirm their clear separation from other elite communities. Second, the pairwise co-appearance frequencies between other communities is often small (less than 13%). However, the following four distinct groups of elite communities frequently co-appear together (>88% of the time): *(i)* US-Corp, US-TV and TH, *(ii)* ID and K-PoP, *(iii)* IN and PH, and *(iv)* US-Pop and GLB/Fun. We also examine the frequency of co-appearance of elite communities and individual unstable nodes and observe that each unstable node primarily appears with elite communities in one of the above groups [14]. This suggests that *individual unstable nodes act as "hubs" and facilitate tighter coupling between the corresponding elite communities*. The low frequency of co-appearance between two elite communities suggests that we can consider them as rather unrelated components of the elite network. Therefore, we can conclude that *the* 10K-ELITE *view of the Twitter elite network consists of 10 separate components, the above four groups and six individual elite communities.*

V. 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 (*i.e.*, mostly regular) 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 engagement and user degree are generally correlated [9], the ranking of influential users based on user engagement and user connectivity measures (*e.g.*, PageRank) are not strongly correlated [11], [7]. There are four important differences between our analysis of influence between elite communities and prior studies [7], [11], [3], [8]. 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 *retweets* and *replies* to quantify pairwise influence between elite users. Third, we characterize cross influence at the granularity of elite communities as well as individual users. Fourth, we examine the relationship between community level influence and community level importance 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 an elite user u (in terms of retweet or reply) on all other elite users with two metrics, namely *(i) number of influenced elites*: the number of unique elite users who have retweeted (or replied to) at least one of u's original tweets (an indication of how widespread u's influence is); and *(ii) aggregate influence*: the summation of the fractions of any other elite user's captured tweets that are retweet of (or reply to) tweets originally



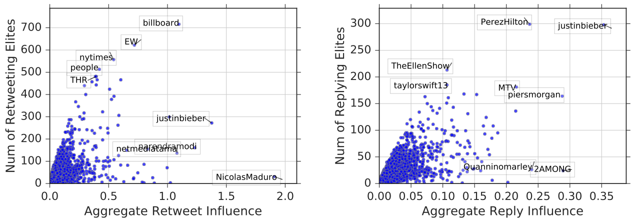
(a) Num. of Retweets (b) Retweet Influence (%)

Fig. 5. Directed pairwise community-level influence

generated by u (an indication of the aggregate magnitude of u's influence for this user). More specifically, this aggregate influence for user u is defined as $AggUserInfl(u) = \sum_{v \in Elite} \frac{RT_{u \to v}}{N_v}$ where $RT_{u \to v}$ 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 for any user in C_i on any user in C_j as $AggCommInfl(C_i, C_j) = \sum_{v \in C_i} \sum_{w \in C_j} \frac{RT_{v \to w}}{N_w}$ To conduct this analysis, we collect all available tweets of all accounts in 10K-ELITE. Our dataset contains more than 31M tweets where 6.5M of them are retweets and 5M are replies.

Community-Level Influence: In Figure 5(a) the color of cell (i,j) indicates the absolute number of times that a user in elite community i has retweeted tweets originated by users in elite community j. This heatmap shows that members of each elite community primarily influence other members of their own community. Furthermore, larger communities also influence other (smaller) communities. Interestingly, the level of influence is generally balanced between communities. Since these absolute metrics could be biased towards larger communities, Figure 5(b) presents the normalized view of influence where the color of cell (i,j) indicates the percentage of tweets by users in community i that is a retweet of tweets originated by users in elite community j. This normalized view provides a more proper representation of the influence between elite communities. Surprisingly, this measure has non-zero values mostly on the diagonal cells. Furthermore, the level of influence within elite communities is not a function of their size. Elite users in Adult, IN and Arabic have the most retweet influence. We observe very similar result for reply influence (available in [14]) with elites in K-PoP, IN, and TR showing the most reply influence on their community members. These results demonstrate that *both the retweet and reply influence of elite users are primarily contained within their own elite community.* The only noticeable exception to this dominant pattern is the retweet influence of US-Corp on US-TV.

User-Level Influence: To gain more insight, we also characterize the patterns of pairwise user-level influence among elites. Figure 6 depicts both dimensions of influence for individual elite users in a scattered plot where each point presents a user, its x-value indicates the user's aggregate retweet (or reply) influence and its y-value shows the number of unique elite users influenced by that user. In Figure 6(a), on the one extreme, we observe some elite users (*e.g.*, @nicolasmaduro, President of Venezuela) that exhibit large retweet influence but on only a small number of elite users. On the other extreme, some elite accounts, associated with news (*e.g.*, @Billboard and @EW) exhibit a lower aggregate influence but on a larger number of elite users.



(a) Retweet Influence (b) Reply Influence

Fig. 6. Influence of individual elite users on all other elites.

Upon further examination of the reply influence in Figure 6(b), we observe that *(i)* both dimensions of the reply influence exhibit a lower value than the corresponding one for retweet influence, and *(ii)* while the most influential replying elite users are generally celebrities in the entertainment industry and gossip media (*e.g.*, @PerezHilton the gossip blogger and columnist), the most influential retweeting elites are often news agencies and political figures. *This analysis illustrates that both dimensions of retweet or reply influence are equally important to gain meaningful insight into the nature of influence for individual elite accounts.*

Influence vs Importance of Elite Communities: An intriguing question is *whether the relative influence of elite users in a community is related to their relative position in the elite network?* To answer this question, Figure 7 presents the summary distribution of the rank among all elite users in 10K-ELITE based on the two measures of influence (*i.e.*, retweet and reply) for users in each elite community (including unstable nodes) using the blue and purple bar, respectively. Furthermore, we also include the summary distribution of user ranks based on the user's PageRank [6] in the elite network as an overall measure of importance among users in each elite community (shown as green bars in Figure 7). Note that each one of these summary distribution of ranks for users in an elite community demonstrates a different aspect of their influence. Figure 7 shows that the relative ranking of users based on different influence measures results in comparable ranges for most elite communities. This observation suggests that these different measures of influences are indeed related. However, there are also a few communities (*e.g.*, Arabic, IN, K-PoP and Adult communities) that exhibit very different rankings for various influence measures. For example, users in the IN community have a high reply influence, moderate retweet influence but low importance ranking. To contrast, users in the Arabic community show a pattern with a higher importance ranking, much lower retweet ranking and even lower reply ranking. These patterns basically reflect the nature of the overall influence of an elite community on the rest of elite network.

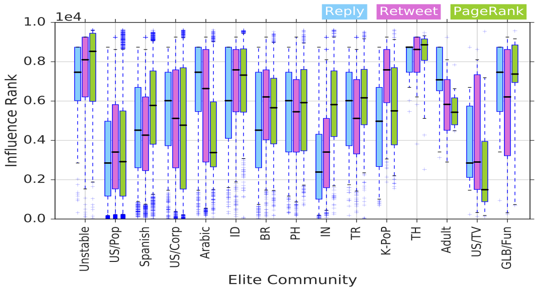


Fig. 7. Distribution of the rank of users in each elite communities in 10K-ELITE based on two measure of influence and PageRank

VI. FROM COMMUNITIES TO PARTITIONS

One of the main benefits of identifying elite communities is to use them to gain more insight into the structure of the Twitter network as a whole. As reported earlier (in Section II), more than 80% of regular (*i.e.*, non-elite) Twitter users follow at least one elite user in 10K-ELITE (*i.e.*, have at least one elite friend). This high visibility of elite users coupled with our observation of the existence of socially-cohesive elite communities raises the question *whether the regular users can be broadly divided into meaningful partitions where each partition contains regular users associated with one elite community?*

For an initial investigation of the association between regular users and elite communities (see [14] for more details), we randomly select 10K regular users as representative samples of all Twitter users. Referring to the fraction of elite friends of a regular user u that are located in elite community c as u's *belonging factor* to community c, we observe that a significant fraction of regular users have more than 70% of their elite friends in a single elite community. Therefore these regular users can be reliably mapped to their corresponding elite community with the largest belonging factor. Finally, to test our hypothesis that the collection of regular users that are mapped to a single elite community can be viewed as a *"shadow partition"* of that elite community, 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 ratio is very similar to the fraction of relationships between elite users that are located in different elite communities. In short, since elite communities can be used to partition regular users, they are capable of providing a meaningful macroscopic view of the entire Twitter structure.

VII. CONCLUSION & OUTLOOK

In this paper, we present a socially-informed characterization of the Twitter elite network. We devise a new technique for efficiently capturing the Twitter elite network that contains the top-10K most-followed accounts and their friend-follower relationships. After annotating each node in the resulting network with its social attributes, we identify resilient elite communities in the Twitter elite network and show that they exhibit social cohesion with a clear theme and therefore represent socially meaningful entities of the network. We then characterize both the connectivity and influence among elite communities and show that grouping regular users based on their association with an elite community results in "shadow partitions" whose inter-connectivity mirrors that of the corresponding elite communities. Thus, these shadow partitions can be viewed as extensions of the elite communities across the entire graph, *i.e.*, the elite communities represent a socially-meaningful coarse-grained view of both the Twitter elite network and the Twitter network as a whole.

Our future plans to extend this work include an in-depth study of the temporal evolution of the Twitter network at the level of elite communities (including their associated social themes). We also 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.

REFERENCES

[1] M. Al-Garadi. Analysis of Online Social Network Connections for Identification of Influential Users: Survey and Open Research Issues. *ACM Computing Surveys*, 51, 2018.

[2] 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.

[3] 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.

[4] 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.

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

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

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

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

[11] 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.

[12] 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.

[13] R. Motamedi, R. Rejaie, D. Lowd, W. Willinger, and R. Gonzalez. Inferring coarse views of connectivity in very large graphs. In *Proc. of COSN*, 2014.

[14] R. Motamedi, S. Rezayi, R. Rejaie, R. Light, and W. Willinger. Characterizing twitter elite communities. Tech. Report CIS-2016-15, www.cs.uoregon.edu/Reports/TR-2016-015.pdf, Univ. of Oregon, 2016.

[15] T. Puranik and L. Narayanan. Community detection in evolving networks. In *Proc. of ASONAM*, 2017.

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

[17] 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.

[18] 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*, 17(2), 2009.

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