Dissecting Twitter Elite Network

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

Walter Willinger
Niksun, Inc.
wwillinger@niksun.com

ABSTRACT

Highest degree nodes in Online Social Networks (OSNs) such as Twitter can be viewed as “social elites” or “connectivity hubs” as they are followed by many users and therefore can influence them. All these elites along with their pairwise connections form a structure that is known as the elite network. The elite network serves as the backbone of the OSN structure and thus its characteristics offer valuable insights about the core of the network. Despite their importance, the characterization of elite networks has received little attention among researchers.

This paper presents a detailed analysis of the macro- and micro-level structure of the Twitter elite network at different sizes. We show that the Twitter elite network has a similar “star-shaped” structure at different sizes where 90% of its nodes form the largest strongly connected component (LSCC) in the center. Furthermore, the LSCC is composed of a number of “resilient communities” that exhibit strong social cohesion. The examination of pairwise tightness between these communities reveals the coarse structure (and level of interest) among these communities. Exploring the aggregate influence of individual elites on the rest of the elite network based on three different measures, we demonstrate that each measure identifies different types of influential users.

Keywords
Online Social Networks, Elites, Communities, Social Cohesion, Influence

1. INTRODUCTION

During the past decade, the increasing popularity of major Online Social Networks (OSNs) has led to a growing interest among social and computer scientists in characterizing different aspects of these networked systems [11, 12]. A significant fraction of these studies examine the connectivity structure of major OSNs (e.g., Twitter [12], Facebook[18], Google+[11]). Due to the huge number of nodes and edges in the structure of a major OSN coupled with commonly-reported skewed distribution of node degree, most characteristics (e.g., node level distributions and identified communities) are dominated by low-degree nodes that often play an insignificant social or connectivity role. A small percent-
In Section 3, we explore the macrostructure of the Twitter elite network. In particular, we show that the elite network regardless of its size has a “star-shaped” structure where 90%+ fraction of nodes are located in the largest strongly connected component (LSCC) in the center and the rest of the elites form individual (or small) SCCs whose followers are in the LSCC. Furthermore, the elite network has an “onion-like” layered structure where nodes with the largest PageRank are at the core and other nodes with lower PageRank form layers around the core accordingly. Most directed edges are from nodes in the core (or inner layers) towards nodes in outer layers.

We then focus on the structure of the LSCC in Section 4. Toward this end, we first motivate and define the notion of stable or resilient community and then identify such communities (and their associated nodes) among elites as well as unstable nodes that do not map to any community for all four views of the elite network. Leveraging the social and country attributes of individual elites, we demonstrate that the identified communities exhibit a strong social cohesion (i.e., clear social theme). This in turn confirms that these communities represent meaningful units of the elite network. We characterize relative connectivity and tightness between these communities which in turn reveals inter-community structure (or inter-community interest) in the LSCC. Furthermore, we demonstrate that unstable elites act as “hubs” among two or more communities and also identify community members who serve as an incoming or outgoing bridge to other communities.

Finally, we investigate the influence of individual elites on the rest of the elite network in Section 5 using three different measures, namely PageRank, retweet and reply. We argue that the number of tweets and replies are not necessarily enough to capture influence. We determine the aggregate influence of individual elites on the rest of the elite network and show what factors affect influence based on each measure. We then identify the top-N most influential (groups of) users based on each measure and examine the overlap among them as a function of N. To our knowledge, this paper offers the first macro- and micro-level analysis of connectivity and social characteristics of the Twitter elite network. Section 6 summarizes the main related work and Section 7 concludes the paper.

2. CAPTURING ELITE NETWORK

Our goal is to capture the Twitter elite network - that is a subgraph of Twitter that includes the top N most-followed nodes and and the (follower-following) edges among them. Furthermore, we need to annotate each node with its social and location attributes in order to explore the impact of these attributes in relationship among elites. There are a few issues that we need to address before we achieve this goal as follows. First, we need to specify the minimum number of followers that qualifies a user as an elite. Second, we need to efficiently identify qualified Twitter accounts, their attributes and the connections among them. In particular, since these users have millions of followers, it is prohibitively expensive to find all their pairwise connections by collecting and examining all their followers.

To cope with these challenges, our data collection strategy for capturing Twitter elite network consists of the following four steps:

1. Capturing a list of most-followed Twitter accounts and their attributes.
2. Inferring their pairwise connections.
3. Identifying missing accounts and validating the information.
4. Collecting all profile information and available tweets of discovered accounts.

The details of individual steps are as follows:

*Step 1:* To bootstrap this process, we crawl lists of most-followed accounts from online resources. In particular, marketing web sites such as socialbakers.com, offer professionally maintained list of most followed Twitter accounts in different social categories (e.g., celebrities, actors, sport, community, ...). Each list 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 8 offered categories and create a master list that is the union of all the unique discovered users with their number of followers (and associated rank), their category and location. The resulting master list consists of 59,832 unique users whose number of followers varies from 263-72M, and they are associated with 123 categories and 191 unique countries. We focus on the top 5K accounts (and their attributes) with the most followers from our master list.

*Step 2:* To collect all the connections among the identified accounts in the previous step, our key observation is that the number of friends for elites are often several orders of magnitude smaller than the number of followers. Therefore, instead of followers, we collect the list of friends for each selected account from Twitter (using its API). This implies that the connection between account x and its follower account y (denoted as x → y) is discovered when we collect the friend list of account y, i.e., each edge is discovered from the follower side. This simple observation ensures that all pairwise connections among selected accounts are identified efficiently (without collecting all followers of all accounts). The total number of friends for all elite accounts is 170M that consists of 57M unique users.

*Step 3:* At this point, we have a snapshot of the most-followed Twitter accounts and their pairwise directed connections. We take a few steps to verify whether the collected information is complete and correct. It is indeed possible that the identified top 5K accounts from online sources do not accurately capture the top 5K accounts on Twitter, i.e., some accounts might be missing. Our observation is that any such missing account should be among the friends of the identified top 5K accounts. Towards this end, we obtain the number of fol-
followers for all the 10K collected friends that have the largest number of followers in the initial list. If the number of followers for any of these accounts is larger than the minimum number of followers in our initial top 5K list, we add them to the master list (at the proper rank) and identify the edges between these accounts nodes and other top 5K accounts by collecting their friend list. Using this technique, we detected 13 accounts that are between the rank of 1,027 and 4,721 among the top 5K accounts. The small percentage of missing accounts along with their relatively low ranking indicate that our master list is rather accurate.

**Step 4:** We collect all the profile information and all available tweets for the top 5K Twitter accounts. The profile information enables us to verify the provided information by socialbakers, namely the number of followers, category and location. The available tweets ¹ 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 accounts 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 5K-E-ELITE offers a representative view of 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 42.2M to 71.8M followers while the last 10 accounts in the top 5K have around 840K followers and the median number of followers among the top 5K is 1.4M. Therefore, the popularity (and thus importance) of any new account beyond top 5K would be much smaller. Second, Avin et al. [4] have argued that the size of elites in an OSN is roughly the square root of the total population. The estimated population of Twitter accounts is 300M [2] where a fraction of these accounts are abandoned or not active [21]. This suggests that the number of Twitter elites is around 17K. Third, while it is feasible to expand the size of the elite network beyond 5K, reliably collecting the desired attributes (social and location) for these users is very expensive. Therefore, we limit our elite network to 5K. However, to examine whether and how the criteria for selecting elites and thus the size of the resulting elite network affects its structural properties, we consider the Twitter elite network at four different size (or views) that contain top 500, top 1K, top 2K and top 5K and refer to them as 500-E-ELITE, 1K-E-ELITE, 2K-E-ELITE and 5K-E-ELITE, respectively.

2.1 Basic Characteristics

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 the left side of Table 1, including number of nodes and directed and undirected edges (|V|, |E|, |E_d|), reciprocity (Rcp), diameter (Diam), average pairwise distance (APD). We also include the number of connected components and strongly connected components and provided the same metrics for the largest SCC (LSCC) as well on the right hand side of the table. This table clearly shows that as the size of the elite network increases (from 500 to 5000), it becomes denser, the fraction of reciprocated edges drops, its diameter (average path length) initially decreases (increases) and then stabilizes. Interestingly, we observe that except for 500-E-ELITE, all other views of the elite network have a single connected component. However, the number of strongly connected components (SCC) grows roughly proportional with the size of the elite network. Interestingly, APD and diameter is comparable for all networks. Another interesting point is reciprocity (32-40%) which is much higher than the reported value for the entire Twitter graph (22%) [12].

The rank correlation between the number of public vs elite followers for top-5K elite is around 0.55 while between their public vs elite friends is 0.1, i.e., popularity of elites among all users and elites are moderately correlated.

3. MACRO-LEVEL STRUCTURE

In this section, we conduct strongly connected component analysis on the elite networks in order to reveal their macro level 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 few cases a handful of nodes). Table 1 summarizes the key characteristics of the LSCC in each view of the elite network. This table shows that the LSCC in each view contains 89-94% of all nodes and 92-96% of all edges of the elite network. Comparing each view of the elite network with its corresponding LSCC, we observe that the LSCC includes a slightly larger fraction of reciprocated edges, slightly shorter diameter and average path length.

To gain more insight into the structure of the elite network, Figure 1 visualizes each view as a directed graph where the LSCC is shown as a green circle in the center and nodes of all other SCCs along with their edges are shown individually. 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. In a few cases, we observe directed edges between nodes outside the LSCC. We recall the direction of edges are from a friend to a follower (or the direction that tweets are propagated. Therefore, Figure 1 indicates that nodes in the LSCC have interest in and receive tweets from nodes in other SCCs (through the elite network) but the opposite is not true. Most outsider nodes are in a SCC with a single node and few of them consist of two or more nodes with a circular connectivity. For example the Pope has four accounts that only follow

¹Twitter only provides the last 3,200 generated tweets by each user.
### Table 1: Basic characteristics of the elite networks and their LSCC

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>LSCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eu</td>
<td>Rep Daim APD No. CC No. SCC</td>
<td>Rep Daim APD</td>
</tr>
<tr>
<td>500-ELITE</td>
<td>500</td>
<td>17,411</td>
</tr>
<tr>
<td>1K-ELITE</td>
<td>1000</td>
<td>48,424</td>
</tr>
<tr>
<td>2K-ELITE</td>
<td>2000</td>
<td>135,595</td>
</tr>
<tr>
<td>5K-ELITE</td>
<td>5000</td>
<td>503,320</td>
</tr>
</tbody>
</table>

*Figure 1:* The connectivity of strongly connected components of the elite networks.

*Figure 2:* The dynamics of LSCC as the network expands each other but they are followed by many accounts inside the LSCC.

Note that as more nodes are included in the view of the elite network, they may act as a shortcut and pull an outsider node in one view into the LSCC in the next view. Figure 2 illustrates how the outsider nodes in each view are mapped/split to the LSCC and outsider in the next view, using a Sankey diagram[1]. Examination of these views shows that roughly 13 – 20% of outsider nodes are pulled into the LSCC in the next view. A subset of outsider nodes in each view have no friends (i.e., no incoming edges) and thus remain as an outsider in all the larger views of the elite network.

Another interesting question is “whether the centrality of individual elites in the elite network is a function of their overall popularity (i.e., total number of followers)?” Given the directed nature of the elite network, we use the PageRank [19] of each node as a measure of its centrality. Figure 3(a)

---

3 An interactive visualization of this diagram is available on our project page

---

4. **ZOOMING INTO LSCC**

In this section, we characterize the connectivity structure within the LSCC.

4.1 **Detecting Communities Among Elites**

Since page rank depends on the graph size, we need to compare the page rank of all nodes in the largest view of the elite network.
We start by exploring whether there are groups of tightly connected nodes (or communities) inside the LSCC and then use these communities as the basic elements to explore the connectivity structure inside the LSCC in each view of the elite network. We emphasize that the identified communities in (different views of) the elite network could be very different from communities on the entire Twitter structure that contain many non-elite (i.e., unpopular) users.

There are two basic issues in identifying communities in the elite networks. First, most commonly-used community detection techniques take undirected graphs as input while the elite network is a directed graph. To address this issue, we convert each view of the elite network into an undirected graph (i.e., turn individual or pair of reciprocated directed edges into a single undirected edge) before using a community detection technique similar to prior studies (e.g., [14]). Second, the outcome of the most commonly used community detection techniques (e.g., Louvain [7], BigCalmm [27], InfoMap [22]) is non-deterministic. More specifically, multiple runs of a single algorithm on the same graph produces different number of communities and/or different mapping of nodes to communities. We use COMBO [23] that relies on multi-objective optimization to find tight communities. Our stability comparison of well known community detection algorithms showed that COMBO results in the most stable communities while maximizing the community modularity. To minimize the effect of non-deterministic variations in the identified communities, we adopt the following strategy: We run the community detection technique on each view of the elite network \( n \) times and determine the communities that individual nodes are mapped to in each run in a vector with \( n \) values, called ‘community vector’. Then, we group all the nodes that are consistently (i.e., all \( n \) times) mapped to the same community (i.e., have the same community vector) and refer to such a group as a Stable Element. Clearly increasing \( n \) is more restrictive which may lead to smaller stable elements since more runs can simply split an element to two (or more) smaller ones. We conservatively consider \( n = 100 \) in our analysis. This process also results in groups of nodes for which no other node has the same community vector. We group to this last set of nodes and nodes in stable elements with a size smaller than 10 and refer to them as Unstable Element. In a later section, we describe how closely related elements can be grouped together. Since the term “community” is the well accepted term to refer a group of nodes that are tightly connected, we simply use community to refer to a Stable Element, and when necessary use ‘traditional community’ to refer to a community identified in a single run of the community detection algorithm.

Using this approach, we identified 5, 11, 15, and 19 communities with 10+ nodes that collectively cover 89.4%, 92.5%, 93.6% and 90.2% of all nodes in the LSCC for 500-ELITE, 1K-ELITE, 2K-ELITE and 5K-ELITE, respectively. Closer comparison of these resilient communities with the identified traditional communities from one run of COMBO shows that for similar sizes, the conductance values are very similar. Due to limited space, these results are not reported but can be found in our related technical report [17].

### 4.2 Social/Geo Footprint of Communities

One important question is “whether nodes in a community exhibit any social cohesion”. Answering this question reveals whether the identified community represents a meaningful elements of the elite network or not. To tackle this question, we leverage the category and location information that we collected for elite nodes. We present two histograms that show the number of nodes across 10 most common categories (i.e., social footprint) and countries (i.e., geo footprint) in each community. Figure 4 depicts the social and geo footprint for the eight largest communities in 5K-ELITE. The social and geo footprints of these communities clearly demonstrate that they exhibit significant social and/or geo (i.e., language) cohesion. Since most of these accounts belong to easily recognizable individuals/entities, we can also examine the identity of accounts in each community and use their social context to learn more about the “theme” associated with each community.

Here we describe the dominant theme in a few of the largest communities in 5K-ELITE:

- **C1 - Spanish Speaking (965):** 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. This community consists of numerous Spanish-speaking actors and pop stars, such as the Columbian singer Shakira and Puerto Rican singer Ricky Martin, but also globally popular football icons, such as Cristiano Ronaldo and James Rodriguez, and sports organizations, such as FIFA and the Olympics.

- **C2 - US Pop Stars (729):** This community is associated with celebrities, pop stars and entertainment media. The vast majority of these elites are from the US with the remainder almost exclusively from English-speaking countries. US pop stars, such as Katy Perry and Kelly Clarkson, 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 former Disney stars, such as Christina Aguilera, Britney Spears, and Demi Lovato.

- **C3 - US Corporate Celebrities & Media (712):** This community is associated with the US and Global media stars and corporate elites. Users are mostly associated with the

The social and geo footprints for communities in other views of the elite network are available on our related technical report. Furthermore, the identity of accounts mapped to individual community in each view are available online at [http://onrg.cs.uoregon.edu/elite](http://onrg.cs.uoregon.edu/elite) for readers.
US and UK. This community consists of elite media, corporate and/or global entities. For example, this community consists of a global news and media organizations, such as BBC, Guardian (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, NASA. We refer to this community as “US/Corp”.

$C_4$ - **Turkish & Arab** (466): This community mainly consists of Turkish and Arab elites. Its other interesting trend is that accounts mostly belong to “communities”. Popular Turkish organizations, such as the football club Galatasaray and NTV television networks, are in this community. Among famous Arab accounts in this community, we can mention Al-Arabiya news group and Lebanese singer Nancy Ajram.

$C_5$ - **Indian & South Asian** (367): 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. The community also consists of a small group of Korean and Japanese accounts, such as Korean actor Siwon Choi.

$C_6$ - **Brazilian** (247): Referred to as “BR”, this community is almost entirely populated by Brazilian cultural elite individuals and organizations, such as the football stars, Kaka and Neymar, and the television network, Rede Globo.

$C_7$ - **Indonesian and Malaysian** (231): User in this community are mostly from Indonesia and Malaysia. The type of the accounts are however much more divers with more focus on celebrities and communities. Example of users in this community is Agnes Monica, the Indonesian pop star. We refer to this community as “ID”.

$C_8$ - **Global Communities** (223): Accounts in this community are mostly “global communities”. Looking though these accounts, we see popular fun & humor accounts, such as @girlposts, @ComedyTruth, @Flirtationship, @LovePhrase. We refer to this community as “GLB/Fun”.

$C_9$ - **Fashion Media & Brands & Celebrities** (121): The most dominant focus in this community is fashion media and brands. Noteworthy accounts here include Vogue, People, and Elle magazines. There is also a large number of clothing brands, such as, Victoria Secret, H&M, Dior, and Louis Vuitton. We refer to this community as “US/Fashion”.

$C_{10}$ - **Filipino** (109): Referred to as “PH”, Most accounts in this community are celebrities from Philippines.

The rest of the communities are smaller than 100 nodes. We use a self explanatory name to refer to them and also note their size in in parenthesis, as follows: $US/Sport1$ (89); $US/WWE$ (72) with users closely related to World Wrestling Entertainment; $US/NBA$ (28); $US/Singer$ (19); $ES-US/Sport$ (15); $US/E!/Sport$ (15); $US/Celeb$ (13); $US/Sport2$ (11); $US/Brand$ (11).
In summary, the analysis of social, community footprints along with the identity of accounts in each community reveals a significant cohesion among nodes in each community around themes that are obvious in some cases (e.g., language, region, country, business) and more subtle in other cases (e.g., global communities). The observed cohesion in these communities indicates that they represent meaningful elements of the elite network.

4.3 Communities in Different Views

We separately identified communities in the LSCC for each view of the elite network. Since different views of the elite network are related and merely formed by adding more elite nodes, this raises the following question: "Whether and how communities in different views of the elite network are related?"

To answer this question, we keep track of the communities’ individual nodes in each view. This in turn reveals the overlapping users between two communities in consecutive views and shows the similarity of two communities in different views. Figure 5 shows the relationships among communities in consecutive views as we expand the size of the elite network using a Sankey flow diagram. Each group of vertically aligned boxes represent communities in each view of the elite network that are ordered from the smallest (in the left) to the largest (in the right) view of the elite network. Each edge/flow (from left to right) between two communities indicates the number of overlapping nodes between them. Note that at each view, at least 50% of the nodes are new compared to the previous view. The social and geo footprint of related communities in different views show that the theme of most communities remains the same across different views (e.g., “E1K-5-ES, GB/celeb” to “E2K-5-ES,GB/celeb”) while the theme for some other communities slightly evolves as more nodes join the community or two communities merge together (e.g., “E500-1-US/celeb” evolves to “E1K-1-US,GB/celeb” for which more ‘British’ users joined the community and slightly changed it geo footprint). We can also observe few splits and mergers in the graph (e.g., “E2K-2-IN,TR/celeb” splits into “E5K-6-IN” and “E5K-5-TR Arab”, and “E2K-5-ES,GB/celeb” and “E2K-3-MX,CO/celeb” merge to form a large mostly Spanish speaking community.)

Of course, the evolving theme for some communities are due to the arrival of many new nodes in each view of the elite network. To verify this issue, we expanded the views of the elite network by 20% in each step and observed that any change in the theme of individual community occurs very slowly. In summary, as we expand the size of the elite network, the number of communities could change. However, there is a clear relationship among communities in different views as these related communities represent the evolving view of a group of nodes with a specific and slowly changing social theme. For the rest of our analysis in this section, we primarily focus on communities in the largest elite network (i.e., 5K-ELITE) as it contains the largest number of communities.

4.4 Inter-Community Connectivity

Since communities represent meaningful elements of the LSCC, we can characterize the structure of LSCC at a coarse view with regard to communities and their inter-connectivity. To this end, we explore their pairwise connectivity in terms of (i) direct follow relations, and (ii) indirect pairwise closeness.

Direct Follow Relations: Figure 6(a) sketches the inter-community connectivity in the 5K-ELITE view of the elite network. In this figure, each node represents a community (or a group of unstable nodes) where its size indicates the population of nodes in the community. The directed line from $C_i$ to $C_j$ represents relationship between accounts in $C_i$ and their followers in $C_j$. The width of each line encodes the absolute number of follow relations while its color (level of darkness) encodes the level of bias in connections between two communities. Here the bias is measured by comparing the number of follow relations between two communities with the number in the randomized version of the graph. Figure 6(a) only shows the top 10% most biased inter-community (bias of 0.05 or larger). The remaining 90% of inter-community edges have lower bias while 60% of all edges have a negative bias (i.e., two communities have less connection than the random version of the corresponding graph).

Figure 6(a) shows the communities in 5K-ELITE can be broadly divided into three groups: First, a number of communities with international users (e.g., BR, ID, PH) do not have any significant connectivity to the rest of the elite network. Second, the Spanish, IN and ES-US-Sport communities form a chain but they are separated from the rest of the elites. Third, all the US-based communities form a separate group along with the unstable nodes. We clearly observe a structure among these US-based communities which indicates their relative interest/attention between them. For example, the Fashion community only follows Corporate community and US/star community only has ties with Pop-stars while US/Singer have ties to both US/Corp and US/Pop communities (with higher than random bias).

Indirect Pairwise Tightness: Direct connection between communities is only one aspect of connectivity. Pairwise tightness between two communities is a more subtle measure that illustrates how tightly connected (or close) two communities are. To assess this notion of tightness between each pair of communities in 5K-ELITE, we examine different runs of the community detection technique to determine whether two communities “co-appeared in the same traditional community”. We use the frequency of co-appearance for each pair of communities across all runs as a relative measure of tightness. We recall that as we increase the num-

---

1 Interactive version of this figure is available on our project page.
2 This visualization also available online.
ber of runs of the community detection technique (or \(n\) in Section 4.1) to identify community, it is more likely that a tight community is split into two or more communities. This analysis illustrates whether some groups of communities are tightly coupled and could address any side effect of using large \(n\).

Figure 6(b) summarizes the pairwise tightness between communities in 5K-ELITE where each node represents a community and the thickness of the (undirected) edge between them shows their co-appearence (i.e., indirect tightness). This graph demonstrates coupling between communities, some of which are aligned with the discovered relationships in Figure 6(a). However, some of the relationships in Figure 6(b) are indeed new. For instance, while the figure shows high coupling among US-based communities, it also presents considerable ties between Indian and US/WWE and GLB/fun communities. We believe that these community-level views of the connectivity structure for LSCC clearly demonstrate the level of interest (and thus influence) among these communities in the elite network.

### 4.5 Node Level Analysis

We examine most visible elites to determine what role they play in the elite network by focusing on the LSCC of 5K-ELITE. Toward this end, we consider unstable nodes of the LSCC (5.3% of nodes that were not part of any community) separate from the remaining nodes in the LSCC.

**Unstable Nodes:** We consider 25 unstable nodes with the highest PageRank and examine their frequency of co-appearence with individual communities across different runs of the community detection technique. Figure 7 presents this information as a heatmap where the color of the cell \((i,j)\) indicates the frequency of co-appearence for node \(i\) with \(C_j\). We emphasize that the sum of the values in each row is not necessarily 100% because a node may co-appear with none, one or more communities in each run of the community detection. Furthermore, we recall that the higher the frequency of co-appearence, the tighter the connectivity (i.e., shorter distance) between a node and a community. In fact, a node that frequently co-appears with one (or more) community could be viewed as a (possibly overlapping) member of those communities that was disconnected from them due to our strict co-appearence requirement in forming communities. Therefore, this analysis illustrates any side effect of our co-appearence constraint \((n=100)\) in forming communities at the user level. Surprisingly, there are a few distinct co-appearence patterns among nodes in Figure 7 as follows: (i) Channing Tatum, YouTube, @rustyrockets, and @MrsOsbourn are closely coupled with US/Pop and have less coupling with other US-based communities aside from US/Sport, US/WWE, and US/Brand. (ii) @Nike and Tiger Woods have very similar co-appearence patterns which ties them mostly to Spanish communities and then US-based communities. (iii) Barack Obama, Lakers, and ESPN’s @JalenRose are very close to US/NBA and US/Star and has a moderate distance from a few other US-based group but never co-appears with US/fashion and US/Singer.

**Bridge Nodes:** Since individual communities present groups of related nodes, one natural question is whether specific node(s) act as bridge to/from other communities?. In general, except for three nodes in 5K-ELITE, all other nodes in identified communities have at least one external incoming or outgoing connection. To tackle this question, we make a distinction between **incoming** \((inBr(s))\) and **outgoing** \((outBr(s))\)
The frequency of communities co-appearance of 5K-ELITE

Figure 6: Graph structure at the community level level

(a) Direct connections between communities of 5K-ELITE

(b) The frequency of communities co-appearance of 5K-ELITE

Figure 7: Co-appearance of 25 unstable accounts with highest PageRank with communities

These outliers clearly act as an incoming/outgoing bridge of each community from/to other communities in the LSCC.

Table 2 presents (screen name for) a pair of accounts in each community that have followers and friends in the largest number of other communities as inBr and outBr for their community, along with the number of external communities that each bridge node is connected with. We observe that most of the outBr are well-recognized and popular individuals/entities with diverse appeal across different groups. However, the inBr are either not recognizable or have genuine interest to many different groups. Given the tight coupling among several US-based groups (as we discussed earlier in subsection 4.4), having followers or friends among these related groups may not necessarily indicate their diversity.

bridge node(s) for community s that have incoming (outgoing) connections from (to) users in the largest number of other unique communities. In other words we measure the importance of a bridge node based on the number of unique communities that it connects to rather than the actual number of connections. For example, if node n1 has 1000 external followers (i.e., followers in other communities) that are all located in two communities and node n2 has 100 external followers that spread across 10 communities, we consider n2 to be an inBr(s) as it is visible among a larger number of communities. Figure 8 presents the summary distribution of the number of unique external communities. In essence, Figures 8(a) and 8(b) illustrate that users in each group of elite generally pay attention (i.e., follow) and receive attention (i.e., being followed) from how many other groups of elites, respectively. We observe that apart from a couple of exceptions (US/E!, Sport and US/Star), most groups typically pay attention to 3 – 7 other groups. This level of attention varies among users in each group but it is often between 1 to 9 groups. Figure 8(b) paints a different picture as it shows that the typical number of unique communities where a user has followers (from which a user receive attention) widely varies among different communities. For example, some communities have followers in 4 to 7 while others have followers in 12 to 15 communities. In both Figure 8(a) and 8(b), there are few outliers in most communities (shown with stars) that have friends and followers in almost all other communities. These outliers clearly act as an incoming/outgoing bridge of each community from/to other communities in the LSCC.

In Figure 7, we illustrate the co-appearance of 25 unstable accounts with highest PageRank with communities. For example, if node n1 has 1000 external followers (i.e., followers in other communities) that are all located in two communities and node n2 has 100 external followers that spread across 10 communities, we consider n2 to be an inBr(s) as it is visible among a larger number of communities.

The summary distribution of the number of unique external communities is presented in Figure 8. In essence, Figures 8(a) and 8(b) illustrate that users in each group of elite generally pay attention (i.e., follow) and receive attention (i.e., being followed) from how many other groups of elites, respectively. We observe that apart from a couple of exceptions (US/E!, Sport and US/Star), most groups typically pay attention to 3 – 7 other groups. This level of attention varies among users in each group but it is often between 1 to 9 groups. Figure 8(b) paints a different picture as it shows that the typical number of unique communities where a user has followers (from which a user receive attention) widely varies among different communities. For example, some communities have followers in 4 to 7 while others have followers in 12 to 15 communities. In both Figure 8(a) and 8(b), there are few outliers in most communities (shown with stars) that have friends and followers in almost all other communities. These outliers clearly act as an incoming/outgoing bridge of each community from/to other communities in the LSCC.


Figure 8: Distribution of number of external communities for nodes in each community in 5K-ELITE

We have examined the page rank of the incoming and outgoing bridge nodes in Table 2 to see whether these bridges are among the most central nodes in the elite network. While a few of the central nodes are among the (mostly outgoing) bridges, most of the bridges do not exhibit high centrality, i.e., high centrality of nodes inside different communities does not necessarily imply their role as a bridge in their community.

5. INFLUENCE ON ELITES

In this section, we turn our attention to how individual elite may influence other elites on Twitter. Influence is a rather subtle effect of user \( u \) on one or more other users that can not be properly measured by a single metric. Therefore, we consider the following three different metrics to measure the influence of elite user \( u \) on other elite in 5K-ELITE: (i) Page Rank: page rank of user \( u \) measures its centrality in the elite graph that shows her relative reachability to other elites which in turn affects how easily a tweet by \( u \) reaches (and influences) other elites (ii) Replies: replies show the reaction of other elites to tweets that are generated by user \( u \). (iii) Retweets: retweets demonstrate how tweets by user \( u \) have been relayed (because of their perceived importance) by other elites. PageRank is solely based on connectivity in the follow graph whereas the other two measures capture pairwise interactions between elites based on their generated tweets. Next, we describe issues associated with identifying most influential elites with respect to each one of these metrics, and then examine the overlap among influential users based on each metric.

Table 2: The out (in) bridge nodes in communities along with the number of communities that are following them (being followed by them)

<table>
<thead>
<tr>
<th>Community</th>
<th>User</th>
<th>( inBr(s) )</th>
<th>User</th>
<th>( outBr(s) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>hootsuite</td>
<td>18</td>
<td>hootsuite</td>
<td>18</td>
</tr>
<tr>
<td>US/pop</td>
<td>MTV</td>
<td>18</td>
<td>TheEllenShow</td>
<td>18</td>
</tr>
<tr>
<td>US/corp</td>
<td>UNICEF</td>
<td>18</td>
<td>billboard</td>
<td>18</td>
</tr>
<tr>
<td>TR/Arab</td>
<td>drangelocarbone</td>
<td>18</td>
<td>lonelyplanet</td>
<td>14</td>
</tr>
<tr>
<td>IN</td>
<td>PleasureEllis</td>
<td>18</td>
<td>PleasureEllis</td>
<td>16</td>
</tr>
<tr>
<td>BR</td>
<td>dawallach</td>
<td>18</td>
<td>giselleofficial</td>
<td>15</td>
</tr>
<tr>
<td>ID</td>
<td>UberSoc</td>
<td>13</td>
<td>OneLouderApps</td>
<td>18</td>
</tr>
<tr>
<td>GLB/fun</td>
<td>Drake</td>
<td>18</td>
<td>Support</td>
<td>17</td>
</tr>
<tr>
<td>US/fashion</td>
<td>DrJimmyStar</td>
<td>17</td>
<td>chelseahandler</td>
<td>18</td>
</tr>
<tr>
<td>PH</td>
<td>jungjungs</td>
<td>16</td>
<td>MannyPacquiao</td>
<td>16</td>
</tr>
<tr>
<td>US/sport</td>
<td>VernonDavis85</td>
<td>18</td>
<td>expn</td>
<td>18</td>
</tr>
<tr>
<td>US/WWE</td>
<td>JonahLupton</td>
<td>18</td>
<td>MensHumor</td>
<td>16</td>
</tr>
<tr>
<td>US/NBA</td>
<td>Baron_Davis</td>
<td>16</td>
<td>NBA</td>
<td>17</td>
</tr>
<tr>
<td>US/singer</td>
<td>eonline</td>
<td>18</td>
<td>eonline</td>
<td>17</td>
</tr>
<tr>
<td>ES-US/sport</td>
<td>verified</td>
<td>18</td>
<td>TwitterSports</td>
<td>18</td>
</tr>
<tr>
<td>US/brand</td>
<td>PopWrapped</td>
<td>18</td>
<td>GMA</td>
<td>18</td>
</tr>
<tr>
<td>US/star</td>
<td>pizzahut</td>
<td>13</td>
<td>SHAQ</td>
<td>17</td>
</tr>
<tr>
<td>US/sport</td>
<td>HEELZiggler</td>
<td>11</td>
<td>Yankees</td>
<td>15</td>
</tr>
<tr>
<td>US/brand</td>
<td>benludis</td>
<td>18</td>
<td>MichelleDBeadle</td>
<td>17</td>
</tr>
</tbody>
</table>

PageRank: A key question is whether (and to what extent) the PageRank of a node is a function of its in- and out-degree? To explore this issue, Figure 9 presents the out-degree (on the x axis), in-degree (on the y axis), and PageRank (as the color of circles) for all elites in 5K-ELITE using a log scale for both axises. Furthermore, the total degree (or number of unique followers and friends) of a node is encoded in the size of its circle. Not surprisingly, this figure demonstrates that the PageRank of nodes is primarily determined by their out-degree (i.e., number of followers) rather than in-degree or total degree. However, it is interesting that some users (Bill Clinton and Pope Francis’s English account, @Pontifex) whose number of followers are an order of magnitude smaller than the maximum number of followers, still achieve some of the highest PageRank.

Repling & Retweeting: Using Twitter API, we collect the last 3,200 (maximum possible) tweets that Twitter provides for individual elites in 5K-ELITE. For each retweet by user \( u_x \), the API indicates the ID of the corresponding original tweet and user \( u_o \) that posted the original tweet. The number of times that user \( u_o \) retweets a tweet that is generated by user \( u_o \) depends on the tweet rate of these two users. For example, if \( u_o \) has a low tweet rate and one of its tweets is retweeted by \( u_x \), this indicates higher influence than the case where \( u_o \) has high tweet rate. On the other hand, if \( u_x \) has a low tweet rate and retweets a tweet from \( u_o \), this indicates a higher influence by \( u_o \) compared to a case where \( u_x \) generates many tweets and one of them happens to be a retweet from \( u_o \). Furthermore, to measure the overall influence of user \( u_o \) on all other elites, we need to properly aggregate its
pairwise influence on all other elites.

To incorporate all these factors, we capture the aggregate influence of user \( u \) (in terms of retweet or reply) on all other elites with the following three numbers: (i) **Average Tweet Rate** is measured as the ratio of \( u \)'s tweet count divided by account age, (ii) **Number of Influenced Elite** is the number of unique elites that have retweeted (or replied) at least one of \( u \)'s original tweets, (iii) **Aggregate Retweet/Reply Influence** of user \( u \) is the sum of the fractions of any other elites' recent tweets that were retweeted or replied to tweets originally generated by \( u \), more specifically \( \text{AggInfluence}(u) = \sum_{j \in \text{Elite}} \frac{RT_{i \rightarrow j}}{N_j} \), where \( RT_{i \rightarrow j} \) and \( N_j \) denote the number of times that user \( j \) retweeted (or replied to) user \( i \) and \( N_j \) is the total number of \( j \)'s tweets. Figure 10 shows this three-dimensional measure of retweet influence for all elites in 5K-ELITE as a scattered plot. To present a more clear picture, in figure 10(a) and 10(b) we remove users that have influence on less than 50 and 25 elite, respectively. Each circle represents an elite user \( u \) where its \( x \) coordinate indicates \( u \)'s average tweet rate (with log scale), and its \( y \) coordinate shows \( u \)'s aggregate retweet influence. We also encode the number of influenced elites for each user in both the size and color of each circle for clarity and also labeled some of the key circles with their account name. Figure 10(a) illustrates a few interesting points: First, we observe quite a few users whose aggregate influence is rather large (> 0.4) despite their low tweet rate but they are being retweeted by a small number of elites. For example, @narendramodi (Prime Minister of India) has aggregate retweet influence of 0.65 but is being retweeted by only 72 elites. In essence, these users have a high retweet influence on a small number of elites. Second, we can easily recognize a number of users who are being retweeted by more than 200 elites. Their aggregate influence widely varies among them and does not seem to be correlated with their tweet rate. Most of these accounts belong to news and media agencies, e.g., @billboard is being retweeted by 400+ elites.

Figure 10(b) presents the three dimensions of reply influence among elites with an encoding similar to Figure 10(a). This figure shows some clear differences between reply and retweet influence. First, the number of replying users and aggregate reply influence are roughly an order of magnitude smaller than the corresponding measure for retweets. Second, there are a number of accounts with a wide range of tweet rates that receive replies from many other elites. Almost all of these belong to individual celebrities and the most active and influential one is Perez Hilton with reply from 220+ elites.

The observed minimal overlap among the top-10 most influential users based on different measures raises the following question: “how does the overlap among the top-N most influential users based on different measure change with N?” Exploring this question reveals the level of separations between the influential users that are identified by each metric. The three Venn diagrams in Figure 11 present pairwise and three-way overlap among top-N influential users based on three categories for N=25, 100 and K. We observe that the three-way overlap among different group of influential users...
Figure 11: Overlap among difference influence measures grows with N from 12% to 21% and 47%. Interestingly, even for top-1K scenario, between 21 – 28% of users are considered influential based on a single measure. Finally the plots do not reveal any similarity between ranking observed by any two metrics, i.e., each metric captures a distinct essence of influence.

6. RELATED WORK

A few prior studies have examined elite networks in major OSNs. Avin et al. [4] reported that the attributes of the elite network are very different from the whole graph and also can not be reproduced by graph generation and affiliation models. They also show that the elite network has a much higher density, reciprocity and average degree compared to the rest of the social network. They concluded that elites make up roughly top N nodes where N=√n and n is the total number of nodes. Wu et al. [26] uses “lists” in Twitter to identify what they call elite users. They define elites as users that frequently appear in lists of certain kinds (celebrities, media, blog, organization) but their final set of elite users is arguably users with the large number of followers. Although they look into the interconnection among elites, they only group them based on just four predefined categories. None of these studies have explored community level analysis and user influence on elite network.

Measuring influence through social media has also been studied in the past. Kwak et al. [12] ranked users by i) their number of followers, ii) PageRank over the follower graph and, iii) their number of retweets to identify the most influential users. They reported a higher correlation between number of followers and PageRank metrics compared to the correlations among other pairs of metrics. Welch et al. [25] calculate PageRank over both retweet graph and follow graph, and then compare the resulting rankings. They conclude that PageRank over the follower graph reveals the popularity of a user and over the retweet graph demonstrates user influence. However, since their retweet graph captures retweet relationship (i.e., has a direct edge from the original sender to each retweeting user), it does not capture the diffusion of the tweet and thus the resulting PageRank may not be informative. Backshy et al. [6] tried to predict the size and depth of a retweet diffusion tree. First, they reconstruct the retweet diffusion tree using a follower graph that is captured more than 10 months prior to the retweet event. This timing gap obviously leads to error in their inferred diffusion trees. Using user attributes and attributes extracted from the inferred diffusion trees, authors try to predict the depth and a size of new diffusion trees. Deng et al. [9] argue that past history is not sufficient when measuring retweet probability. Instead, they assume that users are primarily influenced by their friends. Using the connectivity graph of Weibo (Chinese Twitter) and retweet information, they use a Bayesain method to estimate pairwise influence of users and therefore predict the properties of new diffusion trees. Wu et al. [26] also study the influence on Twitter. However, their main focus is to capture “who listens to whom on twitter” and whether conventional media sources play a different role in message propagation to non-elites compared to other elites. There are two key differences between our work and prior studies on influence. First, we consider multiple measures to capture influence. Second, and most importantly, we focus on cross influence among elites rather than the influence of elites on ordinary users.

7. CONCLUSION

In this paper, we characterized structural properties of the Twitter elite network and revealed its macro-level structure with its LSCC as its core component. We showed that LSCC is composed of tightly connected communities with strong social cohesion and unveil the substructure among these communities. Finally, we assessed the aggregate influence of each elite user on the rest of the network using three different measures and explored the similarity among top-N influential users based on various measures. We plan to extend this work by investigating the temporal evolution of elite network along with its underlying causes and implications.

8. REFERENCES


