Characterising and Evaluating Online Communities from Live Microblogging User Interactions

Hugo Hromic and Conor Hayes
Insight Centre for Data Analytics @ NUI Galway, Ireland
{first.last}@insight-centre.org

***Abstract*—Microblogging (mainly represented by Twitter) is a type of social media that focuses on fast open real-time communication using short messages between users and their followers. This system is attractive due to its open nature and agile content sharing, leading to a compelling and popular social media platform which generates large amounts of content by the minute. Community finding techniques are an interesting approach for organising this massive content but there is no clear agreement in the literature for a standard definition of *user community* for the microblogging use case, leading to unreliable ground-truth data and evaluation. In this work, we differentiate between *functional* and *structural* definitions of communities for microblogging. A functional community groups its users by a common independent social function, e.g. fans of the same football team, while in a structural community the members exclusively depend on their connectivity in a network, e.g. modularity. We build and characterise eight types of functional communities to be used as user-labelled ground-truth and five types of live user interactions networks from Twitter. We then evaluate thirteen popular structural community definitions using five different Twitter datasets, exploring their goodness and robustness for detecting the functional ground-truth under different perturbation strategies. Our results show that definitions based on internal connectivity, e.g. Triangle Participation Ratio, Fraction Over Median Degree or Conductance work best for the Twitter use-case and are very robust. On the other hand, classic scores such as Modularity are limited and do not fit very well due to the sparsity and noise of microblogging. An implementation of our experimental framework is also made available.**

I. Introduction

Online Social Networks (OSN) have developed from simple static blogs into richer interactive systems such as Facebook, Instagram, Twitter, LinkedIn or YouTube. An attractive type of OSN is *microblogging*, which allows for fast real-time open broadcasting of short content among friends and/or followers. Examples of microblogging OSN are Twitter, Weibo (the Chinese counter-part of Twitter), and Tumblr, which is similar to Twitter but focused on multimedia posts. Twitter is currently one of the most widely known microblogging OSN in the world, with more than 330 million monthly active users as of December 2017 [1], generating an average of 500 million *Tweets* (short messages) per day.

While OSN connect us to friends and acquaintances, they also fuel an increasing content disorganisation for its users. This problem can be alleviated with content filtering or clustering techniques such as community detection [2]. However, a fundamental challenge in community detection is the lack of agreement in the literature for a definition of *user community*, which makes them difficult to evaluate and interpret. For the case of microblogging, researchers often use the same definition as for more traditional social media [3]–[6], or definitions from topic analysis and user profiling [7], [8]. However, are these definitions appropriate for microblogging? We argue that these adoptions might not be suitable for microblogging due to its particular fast-pace and user sparsity.

Instead of attempting to craft yet another community definition for microblogging, we will prefer and evaluate a more flexible non-personal wider interpretation. We hypothesize that, in microblogging, people do not seek to be closely related but instead are more curious about the collective opinion of the masses. We will then differentiate between *functional* and *structural* definitions of user communities [9]. A functional community groups its users by a common independent social *function*, e.g. fans of the same football team, while in a structural community the members exclusively depend on their *connectivity in a network*, e.g. their average node degree. We then argue that functional communities can be uncovered from structural patterns in a network of live interactions.

In this work we propose to characterise and investigate the network structure of functional communities considered as user-labelled ground-truth, for the microblogging OSN scenario. The communities under study can emerge from real-time microblogging user interactions such as replies, user mentioning, posts rebroadcasting and quoting. Static structural sources, e.g. followers networks, are commonly used for this purpose, however they are often prohibitive to capture for global analysis and are not suitable for detecting fast-paced community formation and termination [3]. Therefore, we propose to use live streams of interactions instead.

Our main motivation is that it is difficult to identify user communities in live streams of microblogging OSN due to their velocity and low reciprocal characteristics. We hypothesise that the highly dynamic and fast-paced nature of microblogging causes users to switch or loose interest about topics quickly, rendering conventional community discovery based on more static and dense networks less effective.

We address the following fundamental research question: does a distinguishable correlation exist between the underlying network of user interactions and independent user-labelled functional communities in the microblogging use case of Twitter?. Furthermore, we propose the following sub-questions: (1) how can we create reliable independent sets of ground-truth functional communities from Twitter? (2) how can we create reliable graph models to represent live user interactions in Twitter? (3) do structural patterns exist in a network of interactions from Twitter that align to ground-truth functional communities?, and (4) how do existing structural community definitions perform in live interaction networks?

From these questions, we provide the following contributions: (a) a methodology for building ground-truth functional communities for stream-based Twitter datasets, (b) a graph-based model for live Twitter user interactions, (c) an understanding of global and structural properties of microblogging functional communities, (d) recommendations on community scoring functions based on empirical evaluation of these for Twitter user interactions networks, and (e) an implementation of this study available on request.

The rest of this paper is organised as follows: in the next Section we present background and related work, then we establish and examine sets of ground-truth functional communities for Twitter, including their definitions. Next, we study the performance of a range of existing structural community definitions over our ground-truth and propose recommendations for the microblogging case. Finally we present our conclusions and propose future directions.

II. Background and Related Work

A. Characteristics of Microblogging Social Media

OSN can be categorised according to the degree of social functionalities they offer [10]: identity, conversations, sharing, presence, relationships, reputation and groups. For example, FourSquare focuses on physical user presence, Facebook relies on reciprocal friendship relationships, YouTube focuses on video content sharing, and LinkedIn values user identity.

In the case of Twitter, relationships between users are less personal, promoting instead a more open ambient for socializing [11]. Twitter users can post short messages publicly and other users can reply, quote and *retweet* (rebroadcast) them. Moreover, there is a limit of 280 characters per post, prompting users for brevity and clarity in their content.

A user can choose to follow another user for be up-to-date content, often with no approval needed or without requiring to be followed back, however this mechanism has low reciprocity. Only 20% of users follow each other [12], in contrast to other services such as Flickr (70% [13]) or Yahoo! 360 (80% [14]). This low reciprocity suggests that Twitter followers networks might not be adequate for structural community detection. Information in Twitter spreads less than five hops away, shorter than in other known OSN [12], highlighting the strength of microblogging as a medium for rapid information diffusion compared to other OSN focused on verified relationships.

Microblogging also differentiates itself from characteristics of classic human social networks [15]: the distribution of subscribers is not power-law, the degree of separation is shorter and most links between its users are not reciprocated [12]. However, Twitter evidences degrees of homophily: contact between similar people occurs at a higher rate than among dissimilar members, resembling communities [12].

B. User Communities in Microblogging

In OSN, users develop natural groupings from finding other users with similar interests, i.e. homophily [16]. Furthermore, communities allow users to better focus on interesting content.

The definition of a user community for Twitter is generally described in the literature as *"a group of nodes more densely connected to each other than to nodes outside the group"* [9], [11], [16]–[20]. However, communities can be of very different nature and intentions, and often are based on topical subjects or shared interests, i.e. they are functional to the users [9], [17]. Functional communities have been also suggested to require an intermediate social object that connects people together to truly become social, otherwise they loose interest [21].

The community discovery task for microblogging is mostly addressed by means of exploiting static networks, e.g. followers, captured in snapshots [11], [17]–[20]. In this work, instead we aim to understand how user communities can form solely through their public live user interactions represented as a network. Community detection is also approached via a combination of both methods [3]–[5], however we argue that such static networks are expensive to retrieve and maintain fresh in comparison to a stream of messages [3].

C. Study Methodology

We inspire our research methodology on the work of Yang *et al* [9]. The authors empirically study structural community definitions on classical OSN. However, our work differentiates from theirs in that: (a) we extend the original study from traditional OSN to the microblogging case, taking into account the particularities of the platform, and (b) we adapt the original experiments to address the challenges imposed by microblogging, including data volume and its different characteristics.

III. Ground-Truth Communities in Twitter

In this Section we address our first two research questions: how can we create reliable independent sets of ground-truth functional communities and how can we create reliable graph models from live user interactions in Twitter.

We distinguish two independent definitions of user communities: *functional*, based on social function, and *structural*, based on the connectivity in a network. Functional communities will represent our ground-truth data because users themselves explicitly state the social function of their posts, e.g. referencing the same hashtag or mentioning the same celebrity. On the other hand, a structural community is a set of users with a particular connectivity pattern in the underlying live interactions network, e.g. a high edge density.

Our goal is then to investigate the relationship between these two definitions of communities, considering the task of community detection as the recovery of user communities based on a structural definition that later correspond to ground-truth functional communities [9]. In other words, we aim to find an alignment of connectivity patterns in the interactions network to explicitly labelled social functions.

A. Building Live Interactions Networks

In Twitter, posts can be composed using special syntax for providing searchable *#hashtags*, mentioning other users using *@username* anchors, linking to web resources and embedding media files, e.g. pictures or videos. This special syntax, together with replying to posts and retweeting, can be used to form a network of interactions between users [22].

Based on [23], [24], we consider four types of Twitter interactions for building individual networks from a stream of Tweets: mentions, quotes, replies and retweets. A network $G = (V, E)$ is created with a set of vertices V and edges E. Every time a user $u_i \in V$ interacts with another user $u_j \in V$ using any interaction type, we create an edge $(u_i, u_j) \in E$ in the network. For simplicity, we consider G as undirected and unweighted. We also record the time t for each edge to study the time dynamics of these interactions in future work.

We initially considered building separate networks for each interaction type. However a pair-wise network overlap analysis in our data revealed a low value of $\approx 3.82\%$, mostly between the *mention* and *reply* interaction types.

B. Building Ground-Truth Functional Communities

We build ground-truth functional communities from a stream of Tweets where the members explicitly use a common functional social object of a particular type, independent of their underlying interactions. For example, if a set of users $\{u_1, u_2, u_3\}$ use the same hashtag h, then a ground-truth community $C_h = \{u_1, u_2, u_3\}$ is created.

We consider the following social objects for building ground-truth functional communities from Twitter:

1) **Mentions** to group users that mention the same user, e.g. a celebrity in a recent event.
2) **Replies** to group users that reply to the same user, e.g. a controversial commentary discussion.
3) **Quotes** to group users that quote the same user, e.g. provide an opinion over a statement of a politician.
4) **Retweets** to group users that retweet the same user, e.g. a newscaster posting a shocking news.
5) **Countries**, **Cities** and **Places** to group users posting from the same location at different granularities. In Twitter, a *place* is an optional well-known location object that can be embedded in Tweets. Places can contain country and city attributes, hence we can form functional communities based on these three abstractions.
6) **URLs** to group users that share the same web link, e.g. an interesting cooking recipe.

Users might not be fully aware that these social objects can create connections between them in the form of functional communities. Our objective is then to confirm if such connection really exists through their underlying interactions.

Even though some of these functional types are also used to build the interactions network and could be considered as an inherent bias, we note that in the case of community building, these interactions are always used in context with an external factor and not between the interacting users. For example, a ground-truth functional community $C_m = \{u_1, u_2, u_3\}$ mentioning the same user u_m is built, however u_m does not need to be in C_m nor interact with any members in the community. Instead, u_m is considered as an external social object for the members to be connected, similarly to how they would be linked through common hashtags or locations.

We desire for the users in the functional communities to remain connected in the underlying interactions network, therefore we impose two restrictions during construction: (a) each group must have at least three members to facilitate the study of community scoring functions based on triad participation, and (b) each group must be a single connected component in the underlying live interactions network.

Analogous to the interactions network, we also record the time t when each user joined each community for future work.

C. Experimental Ground-Truth Datasets

In this work, we investigate real-world Twitter data streams under different settings and periods of time. The Twitter Streaming API offers two modes for collection: the *filter* and the *sample* endpoints [25]. The first can retrieve streams using defined keywords, geographical coordinates and users to follow, while the latter provides a global-scale unspecified random sampling (estimated to $\approx$ 1-2%) of all Tweets posted.

In total, we collected five streams from Twitter. Using the filter endpoint, we captured two streams for two major world-wide events, one stream of location-based Tweets, and a fourth stream for different TV shows and their audience. To complement our study, we also captured one stream from the sample endpoint. Our collected datasets from Twitter are:

1) POPE2013, captured during the Catholic Pope Conclave event in 2013. Spans for $\approx$ 2 days and contains 460K Tweets and 285K users. The stream listened for event-related hashtags and users to follow.
2) POPE2013-SPL, captured in parallel to POPE2013 using the sample endpoint – 9.9M Tweets and 8.8M users.
3) WORLDCUP2014, captured during the FIFA World Cup event in 2014. Spans for $\approx$ 34 days and contains 27.1M Tweets and 8M users. The stream listened for event-related hashtags and users to follow.
4) RTE2015. RTÉ is the public TV and Radio broadcaster of Ireland. We captured Tweets related to different TV programmes being broadcasted live by RTÉ. Spans for $\approx$ 63 days and contains 2M Tweets and 720K users. The stream listened for event-related hashtags and users to follow related to each TV programme.
5) IRELAND2017, captured using the location filter configured for Ireland during 2017. Spans for $\approx$ 245 days and contains 7.7M Tweet and 1M users.

A summary of the network properties and number of built ground-truth communities can be found in Table I. A total of 6,164,356 communities were built from Twitter. Note how the average user activity and interaction times are very short for the POPE2013 and POPE2013-SPL datasets.

TABLE I

SUMMARY OF BUILT TWITTER GROUND-TRUTH DATASETS. A_u IS THE AVERAGE ACTIVE USER TIME, A_i IS THE AVERAGE INTERACTION TIME AND A_c IS THE AVERAGE COMMUNITY TIME. ALL TIMES ARE IN HOURS. MICROBLOGGING IS NOTED FOR SHORT-LIVED INTERACTIONS.

Dataset	Timespan (Days)	Nodes	Edges	Communities	A_u	A_i	A_c
POPE2013	≈ 2	238,368	303,742	11,580	2.1082	0.6604	18.7042
POPE2013-SPL	≈ 2	6,593,649	6,140,684	5,672,630	4.0370	0.5098	27.1706
WORLDCUP2014	≈ 34	6,932,106	15,854,811	361,559	114.1077	32.3110	334.9153
RTE2015	≈ 63	643,292	1,446,852	56,025	163.6430	84.8706	687.7683
IRELAND2017	≈ 245	1,067,982	2,826,754	62,562	1483.4306	355.7120	3881.9658

IV. EVALUATING COMMUNITY DETECTION IN TWITTER

In this Section we address our third and fourth research questions: do structural patterns exist in a network of interactions from Twitter that align to ground-truth functional communities?, and how do existing structural community definitions perform in live interaction networks?

We start by analysing the feasibility of finding identifiable structural patterns in ground-truth functional communities. Then we evaluate structural community detection methods in the form of community scoring functions. Finally, we assess the goodness, robustness and sensitivity of those scoring functions when applied to our Twitter functional ground-truth.

A. Identifiable Structural Patterns

To provide evidence of distinctive structural patterns in the network, we perform a comparison analysis of ground-truth functional communities and randomly chosen connected nodes with the same path distribution [9]. If such distinctive connectivity patterns exist compared to randomly selected sets of connected nodes, we likely will be able to discover the functional communities based on their network connectivity.

We first define the sets of nodes that we use for this comparative analysis. For every ground-truth community C_i (of any type) in our datasets, we form a corresponding *non-community* $\tilde{C}_i$ from the interactions network based on the following conditions: (1) where possible, $\tilde{C}_i$ must be of the same size than C_i, (2) like every C_i, $\tilde{C}_i$ must be connected, and (3) where possible, users of $\tilde{C}_i$ must have the same distribution of shortest path distances of C_i.

For the microblogging case, the first and third constraints are not always satisfiable. We approach this problem by first computing the χ^2 distance [26] between the shortest path distances histograms of every C_i and of all potential candidates $\tilde{C}_i$. Then, for the first constraint, if it is not possible to find a non-community $\tilde{C}_i$ of the same size for a ground-truth community C_i, we select the closest candidate $\tilde{C}_i$ that has at least 75% of the size of C_i. Likewise, for the third constraint, if an exact match cannot be found, we instead select the closest candidate $\tilde{C}_i$ in descending order or randomly in case of candidates with the same distribution.

We now define the structural properties that we use to compare structural patterns in the interactions network $G = (V, E)$ for both, communities C_i and non-communities $\tilde{C}_i$:

- **Clustering Coefficient (CCF)** measures how likely is a community to form a *small-world* cluster [27].
- **Average Degree (AvgDeg)**, $2|E|/|V|$, is the average node degree of the members of a community [28].
- **Edge Density**, $2|E|/(|V|(|V|-1))$, is the fraction of total edges possible in a community that are present [28].
- **Cohesiveness** is the fraction of total edges possible in a community that are non-bridging [29]. A non-bridge edge is such that when removed, the number of connected components is preserved. This measure captures the intuition of a well and evenly connected community.

TABLE II

RATIO BETWEEN STRUCTURAL PROPERTIES OF GROUND-TRUTH FUNCTIONAL COMMUNITIES AND RANDOMLY CHOSEN NODES WITH SIMILAR SHORTEST PATH DISTRIBUTION FOR THE RTE2015 DATASET.

C. Type	CCF	AvgDeg	Density	Cohesiv
cities	0.9980	1.0827	0.9997	1.3560
countries	0.3679	0.9640	0.9306	0.4207
hashtags	2.1165	1.2562	1.0880	2.0222
mentions	3.7607	1.7791	1.3546	3.1988
places	1.6498	1.0818	1.0315	1.7075
quotes	2.3623	1.3853	1.1481	2.3705
retweets	3.0220	1.5746	1.1864	2.6789
urls	2.6149	1.2950	1.1472	2.4433
Average	**2.1115**	**1.3023**	**1.1108**	**2.0247**

The above properties are computed for every C_i and $\tilde{C}_i$, and then the average ratio r between them is computed for all community types in each dataset. The results for RTE2015, which contains all the considered types of functional communities, are shown in Table II. Our results were similar across datasets. In this example, functional communities have, in average, 2.11 higher clustering coefficient, 1.30 higher average degree, 1.11 higher edge density and 2.02 higher cohesiveness than their respective non-communities. This suggests that ground-truth functional communities in fact have a distinctive structure compared to randomly chosen nodes in the network.

In general, our results show that the ratios for each defined property are $r \geq 1.0$ in the majority of the ground-truth community types and datasets. The *mentions* community type excels with every property having a strong $r \geq 2.0$ between communities C_i and non-communities $\tilde{C}_i$, suggesting that finding functional communities with a third person as functional object is easier than with other objects. Similar is the *hashtags* type, where it is only weak ($r < 1.0$) in the POPE2013 dataset. This is surprising because it suggests that users do not behave like discussion groups, despite this event being highly susceptible to opinions using hashtags. The

RTE2015 and Ireland2017 have strong differentiable hashtags communities ($r \geq 2.0$). For the latter, this is interesting because the capture was only based on location (Ireland), using no particular topic. This result is likely because of the more durable interactions (Table I).

For the cases of the *retweets* and *urls* functional types, the majority of our datasets also exhibit structurally distinguishable functional communities ($r \geq 1.0$). We note that Ireland2017 does not contain any Retweets because Twitter does not deliver them for location-based capturing. This result is consistent with [12], where retweeting and media links are regarded as core activities for news diffusion in Twitter.

The location-based functional types contain few distinguishable communities for WorldCup2014 and RTE2015. Nevertheless, the *countries* type was found to be distinctive ($r \geq 2.0$), suggesting that this abstraction is the most suitable for building functional communities based on location.

Finally, we also note that the *quotes* type is a recent functionality in Twitter, and thus is not captured in datasets older than 2015. Nonetheless, quotes have a strong differentiable structure ($r \geq 2.0$), suggesting that users interact closely around quotes of interest to them.

B. Structural Community Scoring Functions

The goal of community detection is to uncover sets of users in a network with a certain structural pattern. In this context, community scoring functions can be used to quantify how well a set of nodes fit to the desired structure. In this work, we consider thirteen commonly used community scoring functions pre-classified into four families for evaluation.

For a given set of nodes C, the scoring function $f(C)$ measures the quality of C as a structural community in an undirected network $G = (V, E)$. We define n_c and m_c as the number of nodes and edges in the set C, $n = |V|$ and $m = |E|$ as the number of nodes and edges in G, $d(v)$ as the degree of a node $v \in V$, and b_c as the number of edges on the boundary of C, i.e. edges that point outside of C. Using this notation, we now introduce the scoring functions under evaluation:

(I) Class: Only **Internal Connectivity**

- **Density**† is the fraction of total edges possible in C that are actually present. $f(C) = 2m_c/(n_c(n_c - 1))$
- **Edges Inside**† is the number of edges in C. $f(C) = m_c$
- **Avg. Degree**† is the average node degree of C. $f(C) = 2m_c/n_c$
- **Fraction over Median Degree**‡ **(FOMD)** is the fraction of nodes of C that have degree higher than d_m, where d_m is the median degree of all nodes $v \in V$. $f(C) = |\{u : u \in C, |\{(u,v) : v \in C\}| > d_m\}|/n_c$
- **Triangle Participation Ratio**‡ **(TPR)** is the fraction of nodes of C that belong to a triad. A triad is a set of three nodes that are fully connected to each other in C. $f(C) = |\{u : u \in C, u \text{ is in a triad}\}|/n_c$

(II) Class: Only **External Connectivity**

- **Expansion**† quantifies the number of edges per node in the boundary of C. $f(C) = b_c/n_c$

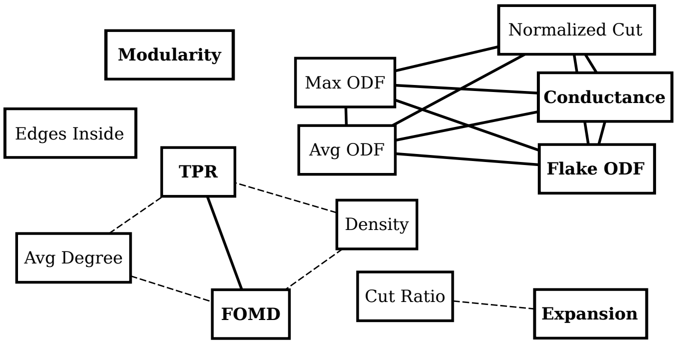


Fig. 1. Groups of scoring functions based on correlation. Weak (≥ 0.3) links are dashed and strong links (≥ 0.6) are solid.

- **Cut Ratio**§ is the fraction of existing edges (out of all possible edges) leaving C. $f(C) = b_c/(n_c(n - n_c))$

(III) Class: **Internal** and **External Connectivity**

- **Conductance**†† is the fraction of total edge volume that is in the boundary of C. $f(C) = b_c/(2m_c + b_c)$
- **Normalized Cut**†† is the cost of cutting edges in C. $f(C) = b_c/(2m_c + b_c) + b_c/(2(m - m_c) + b_c)$
- **Maximum Out Degree Fraction**‡‡ **(ODF)** is the maximum fraction of edges of a node in C that point outside. $f(C) = \max_{u \in C}[|\{(u,v) \in E : v \notin C\}|/d(u)]$
- **Average-ODF**‡‡ is similar to Maximum-ODF but using the average measure instead of the maximum.
- **Flake-ODF**‡‡ is the fraction of nodes in C that have fewer edges pointing inside than to the outside of C. $f(C) = |\{u{:}u{\in}C, |\{(u,v){\in}E{:}v{\in}C\}| < d(u)/2\}|/n_c$

(IV) Class: **Network Model**

- **Modularity**§§ is the difference between the number of edges m_c and the expected number $E(m_c)$ in a random graph with identical degree sequence, i.e. null model.

The details for † can be found in [28], for ‡ in [9], for § in [30], for †† in [31], for ‡‡ in [32] and for §§ in [33].

Preliminarily, we are interested in the relationship between these scores in our Twitter ground-truth datasets. To investigate the contribution of the scoring functions, we first compute each $f(C)$ for each of the six million total ground-truth functional communities C we constructed. Then, we compute a correlation matrix based on the Pearson coefficient and filter it using two thresholds (≥ 0.4 and ≥ 0.6) to unveil connections at different levels between the scoring functions. The result can be seen in Figure 1. With one exception, all of the scores grouped into four clusters, mirroring their pre-defined classes. The *Edges Inside* score remained isolated, even from its close relative *Avg. Degree*. This suggests that, for the case of Twitter, considering only the size of the communities is not enough.

In general, this experiment suggests that despite having numerous structural definitions for communities, they heavily correlate in Twitter. For the remainder of this paper, we will focus on six representative scoring functions from the four classes (bold in Figure 1): FOMD, TPR, Cut Ratio, Conductance, Flake-ODF and Modularity.

C. *Community Detection Goodness*

We now evaluate the community scoring functions in terms of their quality to discover ground-truth functional communities in Twitter streams. In this experiment, we use goodness metrics that capture the notion that good communities should be compact, well connected and well isolated from the rest of the network. The difference between the goodness metrics and the scoring functions under study is that the first quantify a desirable property of the communities, while the latter quantify how community-like is a set of nodes. A community with high goodness does not imply a good scoring function value but a good community score should have a high goodness metric.

We present four goodness metrics $g(C)$. Three of them (Density, Clustering Coefficient and Cohesiveness) were previously introduced as structural properties in Section IV-A. Therefore, a fourth goodness metric is now introduced:

Separability captures the intuition that good communities should be well-distanced from each other [30]. This metric quantifies the ratio between the internal and external edges.

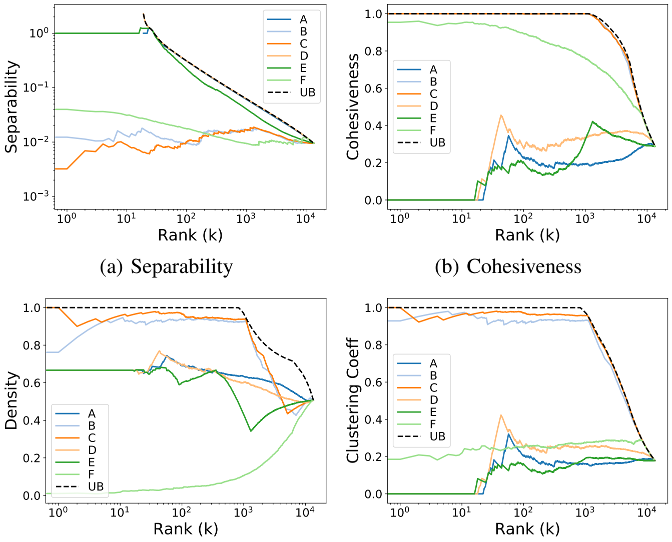
We set up the goodness experiment as follows. For each dataset and community type, we rank our ground-truth functional communities C_i using the six selected scoring functions $f(C_i)$ in descending order. Then, we measure the cumulative moving average (CMA) of each goodness metric $g(C_i)$ for the top-k ground-truth communities under the order induced by $f(C_i)$. A perfect scoring function should rank the ground-truth communities in the same descending order as the goodness metrics, and therefore the CMA should decrease monotonically along k. On the other hand, a poor community scoring function would produce a k-dependent constant CMA.

The results were similar across all of our datasets. For the remainder of this Section we report results for the *hashtags* community type in the IRELAND2017 dataset, representing the rest of the data. Figure 2 shows the four ranked goodness metrics for this representative dataset and community type. The Figure shows the CMA of the six representative scoring functions ranked by the four goodness metrics. An additional upper bound curve (e.g. the CMA of separability ranked by separability) is also provided for reference in the Figure.

We observe that Cut Ratio (A), Conductance (D) and Flake-ODF (E) have a near perfect fit in Separability, while FOMD (B) and TPR (C) show instead an inverse ordering, suggesting that the latter two prefer more dense communities. If the analyst desires denser communities regardless of separation, FOMD and TPR should be preferred. In contrast, for Cohesiveness, Density and Clustering Coefficient, FOMD (B) and TPR (C) prevail with good performance. FOMD and TPR not only prefer denser but also cohesive and packed communities.

Modularity (F) performs relatively well in Cohesiveness. However Cut Ratio (A), Conductance (D) and Flake-ODF (E) exhibit inverse ordering, indicating that these prefer more disperse communities, revealing a failure of the scores to capture cohesive groups in the Twitter scenario.

We observe an interesting near-perfect reversing of Modularity (F) for the Density goodness metric. This is a manifestation of the well-known resolution limit of the Modularity score [34] that becomes evident in our ground-truth functional communities. Modularity (F) also exhibits a near-constant ranking for Clustering Coefficient, suggesting that this score does not prefer or reject well-packed communities.



(a) Separability

(b) Cohesiveness

(c) Density

(d) Clustering Coefficient

Fig. 2. Ranked Scoring Functions for the *hashtags* community type of the IRELAND2017 dataset. Cut Ratio (A), FOMD (B), TPR (C), Conductance (D), Flake-ODF (E), Modularity (F) and their Upper Bound (UB).

TABLE III
AGGREGATED SCORING RANKING BY GOODNESS METRICS.

Score	**CCF**	**Cohesiv**	**Density**	**Separability**
Cut Ratio (A)	5.4302	5.5954	**1.9557**	2.0031
FOMD (B)	2.0270	**1.2837**	3.1018	5.2197
TPR (C)	**1.0289**	1.7561	3.2890	4.8055
Conductance (D)	3.9035	3.9755	3.4466	**1.0069**
Flake-ODF (E)	5.5286	5.2416	3.4173	3.3683
Modularity (F)	3.0774	3.0755	5.7855	4.5948

To complement our results, we also study the ability of the scoring functions to rank the ground-truth communities using the goodness metrics as follows. For each goodness metric $g(C)$ and scoring function $f(C)$, we observe the rank of each score in comparison to the other scoring functions at every rank k. For example, in Figure 2 for Clustering Coefficient at $k = 10^3$, the scores are ranked as: 1st TPR (C), 2nd FOMD (B), 3rd Modularity (F), 4th Conductance (D), 5th Flake-ODF (E) and 6th Cut Ratio (A). Then, for every k we rank and aggregate the six scores using the Borda voting method [35] to obtain an unified ranking that quantifies the ability of each scoring function to find communities with high goodness. The results can be seen in Table III, where ranks near 1.0 indicate good scoring functions for each goodness metric.

Overall, to identify more clustered and cohesive communities in Twitter, FOMD and TOPR are the better choices. If more dense and less separated communities are preferred, then Cut Ratio and Conductance are more adequate.

D. Community Detection Robustness

We now investigate the robustness and sensitivity of the structural community scoring functions in presence of different random perturbations to the ground-truth functional communities. A good community scoring function should be stable under small perturbations and reduce its performance under strong disturbance. The perturbation strategies are [9]:

- **NodeSwap** simulates the effect of community users diffusing from C through the network. First, a random edge $(u, v), u \in C, v \notin C$ is chosen, and then the nodes u and v are swapped. This causes u to abandon C and v to join.
- **Random** perturbs communities by swapping a random member $u \in C$ with a random non-member $v \notin C$. Similar to NodeSwap, this perturbation does not alter the size of the community C but can render it disconnected.
- **Expand** increases the size of communities by choosing random non-members $v \notin C$ that are connected to members $u \in C$, and incorporating them into the community C. This operation decreases the quality of the community.
- **Shrink** decreases the size of communities by choosing random boundary edges $(u, v), u \in C, v \notin C$ and then removing the user u from C. Similar to Expand, this perturbation also preserves the connectedness.

The above perturbation strategies can be controlled using an intensity parameter p, that specifies the number of times $(p|C|)$ the perturbation is applied to a community C.

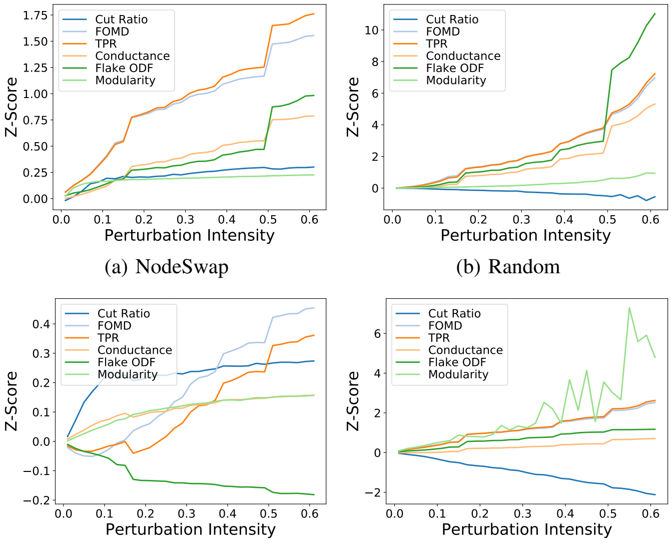
To quantify the impact of applying any perturbation strategy h to a given ground-truth functional community C, lets consider $h(C, p)$ the perturbed version of C under perturbation h with intensity p. Then we measure the Z-score (units of standard deviation) of the difference between the score $f(C)$ of the unperturbed community C and the score $f(h(C, p))$:

$$Z(f, h, p) = \frac{E\left[f(C_i) - f(h(C_i, p))\right]}{\sqrt{Var\left[f(h(C_i, p))\right]}}$$

where $E\left[\cdot\right]$ is the expectancy operator (e.g. the mean) and $Var\left[\cdot\right]$ is the variance operator, both applied over all the ground-truth communities C_i. We note that the sign of TPR, FOMD and Modularity needs to be inverted to ensure that all scores have the same interpretation, i.e. higher is better. Due to the random nature of the proposed perturbations, we repeat them 20 times and average the resulting Z-scores.

The experiment is now as follows: we vary the perturbation intensities $p \in [0.01, 0.60]$ (e.g. in the NodeSwap strategy this means exchanging between 1-60% of the members of a community) and observe the averaged Z-score across all ground-truth functional communities for each type and dataset. The results for the RTE2015 dataset, community type *mentions* are shown in Figure 3. These results are similar for the rest.

The TPR and FOMD scores perform the best in the NodeSwap experiment, followed by Conductance and Flake-ODF. In contrast, Modularity and Cut Ratio do not degrade when we increase the perturbation, revealing their inability to deal with noisy data in Twitter. Similar to the Random strategy, where Flake-ODF takes the lead and TPR/FOMD fall behind. Cut Ratio performs the worst in presence of strong noise.



(a) NodeSwap

(b) Random

(c) Expand

(d) Shrink

Fig. 3. Z-scores for different perturbation strategies applied to ground-truth functional communities in the *mentions* community type of RTE2015.

TABLE IV
AVERAGE ABSOLUTE INCREMENT OF Z-SCORE BETWEEN SMALL AND LARGE COMMUNITY PERTURBATIONS. BEST ARE IN BOLD.

Score	**N.Swap**	**Random**	**Expand**	**Shrink**
Cut Ratio	0.2522	0.1469	0.2605	0.7428
FOMD	**1.6418**	2.7026	**0.4329**	**1.9137**
TPR	0.5866	**3.4041**	0.2658	0.8664
Conductance	0.9777	1.7555	0.1540	0.6294
Flake ODF	1.3473	1.7088	0.3273	0.9439
Modularity	0.4814	1.3332	0.3164	1.4111

The Expand and Shrink strategies also reveal TPR and FOMD as robust scores for Twitter functional communities, and Flake-ODF and Cut Ratio being ineffective in Expand and Shrink respectively. Modularity has consistent good performance in small intensities for Expand and large for Shrink, but degrades with larger expansions and smaller reductions. This is again evidence of its resolution limit [34].

In this experiment, TPR and FOMD are robust functional community scores for Twitter interaction streams, while Modularity and Cut Ratio proved weaker in this context. The Conductance and Flake-ODF scores are a good alternative, however Flake-ODF is not stable for expansion and shrinking.

Finally, we explore how sensitive are the scoring functions in terms of small and large perturbations. We measure the change of Z-score between a small $(p = 0.04)$ and a large $(p = 0.20)$ perturbation, giving preference to scoring functions that quickly degrade in presence of strong perturbations. We averaged the difference $Z(f, h, 0.20) - Z(f, h, 0.04)$ across all our ground-truth functional communities and the results are in Table IV. Large differences indicate that the community scoring function is both robust and sensitive.

In general, FOMD is the most robust and sensitive score in this experiment for all the perturbation strategies except for Random, where TPR takes advantage. Flake-ODF is a close second best for NodeSwap and Expand, Conductance a mild third for Random, and Modularity a close second for Shrink.

V. Conclusions and Future Directions

In this work, we address the problem of evaluating community detection in the context of microblogging services, represented by Twitter. For this, we adopted two interpretations of community: a *functional* definition (used as ground-truth) based on user-labelled social functions, and a *structural* definition based on the connectivity patterns in a network.

We proposed using interactions network and construct explicitly labelled ground-truth functional communities using varied social objects from Twitter streams. Furthermore, we thoroughly evaluated a set of structural community scoring functions from different classes using our ground-truth.

We conclude that scoring functions based on internal connectivity such as the TPR, FOMD and Conductance work best for Twitter, proving to be robust and sensitive. Conversely, the popular Modularity score is limited and unfit due to the sparse and noisy characteristics of microblogging.

More research is required to better understand the nature of microblogging and the community detection task for it. For example, we considered native hashtags as the social function for topics, however other models can be used, e.g. named entities, bag-of-words or TF-IDF. Furthermore, a key social aspect of Twitter is its fast-pacing and openness, pushing for user communities to form and disappear quickly (see Table I) compared to more classic social media such as forums. We need then to scope and shape the concepts of functional and structural communities considering the time-dimension and its significance to their life cycle.

Note: Twitter policies prevent us to share our datasets, however we make available, on request, our framework based on the SNAP engine [36] used to obtain the reported results.

References

[1] S. Aslam, *Twitter by the Numbers (2018): Stats, Demographics & Fun Facts*, en-US, 2018.

[2] S. Papadopoulos *et al.*, "Community detection in Social Media," en, *Data Mining and Knowledge Discovery*, 2011.

[3] D. Darmon *et al.*, "Followers Are Not Enough: A Multifaceted Approach to Community Detection in Online Social Networks," *PLOS ONE*, 2015.

[4] M. Bakillah *et al.*, "Geo-located community detection in Twitter with enhanced fast-greedy optimization of modularity: The case study of typhoon Haiyan," *International Journal of Geographical Information Science*, 2015.

[5] B. Amor *et al.*, *Community detection and role identification in directed networks: Understanding the Twitter network of the care.data debate*. World Scientific, 2016.

[6] N. Cao *et al.*, "SocialHelix: Visual analysis of sentiment divergence in social media," en, *Journal of Visualization*, 2015.

[7] W. Zhou *et al.*, "Community Discovery and Profiling with Social Messages," in *Proceedings of the 18th ACM SIGKDD*, ACM, 2012.

[8] M. Akbari *et al.*, "Leveraging Behavioral Factorization and Prior Knowledge for Community Discovery and Profiling," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, ACM, 2017.

[9] J. Yang *et al.*, "Defining and evaluating network communities based on ground-truth," en, *Knowledge and Information Systems*, 2015.

[10] J. H. Kietzmann *et al.*, "Social media? Get serious! Understanding the functional building blocks of social media," *Business Horizons*, 2011.

[11] D. A. Shamma *et al.*, "Tweet the Debates: Understanding Community Annotation of Uncollected Sources," in *Proceedings of the First SIGMM Workshop on Social Media*, ACM, 2009.

[12] H. Kwak *et al.*, "What is Twitter, a Social Network or a News Media?" In *Proceedings of the 19th International Conference on World Wide Web*, ACM, 2010.

[13] M. Cha *et al.*, "A Measurement-driven Analysis of Information Propagation in the Flickr Social Network," in *Proceedings of the 18th International Conference on World Wide Web*, ACM, 2009.

[14] R. Kumar *et al.*, "Structure and Evolution of Online Social Networks," en, in *Link Mining: MODELS, Algorithms, and Applications*, P. S. Yu *et al.*, Eds., Springer New York, 2010.

[15] M. E. J. Newman *et al.*, "Why social networks are different from other types of networks," *Physical Review E*, 2003.

[16] L. Tang *et al.*, "Community Detection and Mining in Social Media," *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2010.

[17] A. Java *et al.*, "Why We Twitter: Understanding Microblogging Usage and Communities," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007*, ACM, 2007.

[18] A. Gupta *et al.*, "Identifying and Characterizing User Communities on Twitter During Crisis Events," in *Proceedings of the 2012 Workshop on DUBMMSM*, ACM, 2012.

[19] X. Lu *et al.*, "Network Structure and Community Evolution on Twitter: Human Behavior Change in Response to the 2011 Japanese Earthquake and Tsunami," *Scientific Reports*, 2014.

[20] T. Sakaki *et al.*, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," in *Proceedings of the 19th International Conference on World Wide Web*, ACM, 2010.

[21] J. Engeström, *Why some social network services work and others don't—Or: The case for object-centered sociality*, 2005.

[22] H. Hromic *et al.*, "Constructing Twitter Datasets using Signals for Event Detection Evaluation," en, in *Synergies of Case-Based Reasoning and Data Mining Workshop*, 2014.

[23] H. Hromic *et al.*, "Graph-Based Methods for Clustering Topics of Interest in Twitter," en, in *Engineering the Web in the Big Data Era*, Springer, Cham, 2015.

[24] Y. Yang *et al.*, "Automatic Social Circle Detection Using Multi-View Clustering," in *Proceedings of the 23rd ACM CIKM*, ACM, 2014.

[25] F. Morstatter *et al.*, "Is the sample good enough? Comparing data from twitter's streaming API with Twitter's firehose," English (US), AAAI press, 2013.

[26] O. Pele *et al.*, "The Quadratic-Chi Histogram Distance Family," en, in *Computer Vision – ECCV 2010*, Springer, Berlin, Heidelberg, 2010.

[27] D. J. Watts *et al.*, "Collective dynamics of 'small-world' networks," En, *Nature*, 1998.

[28] F. Radicchi *et al.*, "Defining and identifying communities in networks," en, *Proceedings of the National Academy of Sciences of the United States of America*, 2004.

[29] J. Leskovec *et al.*, "Empirical comparison of algorithms for network community detection," in *Proceedings of the 19th international conference on World wide web*, ACM, 2010.

[30] S. Fortunato, "Community detection in graphs," *Physics Reports*, 2010.

[31] J. Shi *et al.*, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[32] G. W. Flake *et al.*, "Efficient Identification of Web Communities," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2000.

[33] M. E. Newman, "Modularity and community structure in networks," eng, *Proceedings of the National Academy of Sciences of the United States of America*, 2006.

[34] S. Fortunato *et al.*, "Resolution limit in community detection," en, *Proceedings of the National Academy of Sciences*, 2007.

[35] D. G. Saari, *Geometry of Voting*, en. Springer Science & Business Media, 2012, Google-Books-ID: bOPwCAAAQBAJ.

[36] J. Leskovec *et al.*, "SNAP Datasets: Stanford Large Network Dataset Collection," 2015.