
Doctoral Thesis

# Methods for Defining Dynamic Online Communities and Community Detection in Fast-paced Social Media Streams

Hugo Fernando Hromic Mayorga

**May 2, 2019**

| External Examiner | Supervisor | Internal Examiner |
|---|---|---|
| Dr. Miriam Fernandez | Dr. Conor Hayes | Dr. John Breslin |


Insight Centre for Data Analytics

College of Engineering and Informatics, National University of Ireland, Galway

# ABSTRACT

Microblogging social media focuses on fast open real-time communication using short messages between users and their followers. Twitter is currently one of the largest and widely known microblogging OSN in the world, with more than 330 million monthly active users as of December 2017. Moreover, an average of 500 million Tweets (short messages) per day are generated within the service. Microblogging social media generate large amounts of content and community finding techniques are a suitable alternative for organising it. However, a fundamental challenge in the community detection literature is the diversity for a definition of user community, which makes evaluating and interpreting algorithms difficult. Therefore, in this thesis, two types of user community definition are adopted and investigated for microblogging: functional and structural definitions. 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.

In this work, functional definitions are built and characterised to be used as user-labelled ground-truth using eight types of social functions from Twitter interaction networks. Afterwards, these ground-truth functional communities are evaluated – in static and dynamic scenarios – considering thirteen popular structural community definitions from the literature. The goodness, robustness and sensitivity of these structural community definitions for detecting the functional ground-truth under different perturbation strategies is investigated. The proposed evaluation is carried using five different Twitter datasets captured during diverse periods of time.

The results of the study show that definitions based on internal and mixed 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, other scores such as Modularity are limited and do not perform well due to the sparsity and noise of microblogging. Furthermore, using user activity as basis to refine communities into their active hotspots further improves the performance of community detection in microblogging. It is demonstrated in this work that standard community detection algorithms are challenged by the fast-paced dynamics and link sparsity of microblogging data. Therefore, it is argued that temporal characteristics must be considered for community detection methods in microblogging.

# CONTENTS

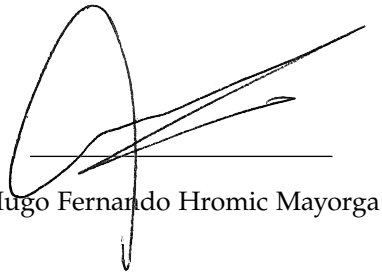
# ACRONYMS

| | |
|---|---|
| AVGDEG | Average Node Degree. |
| CC | Clustering Coefficient. |
| CCDF | Complementary CDF. |
| CDF | Cumulative Distribution Function. |
| CMA | Cumulative Moving Average. |
| CPM | Clique Percolation Method. |
| DCV | Dynamic Community Viewer. |
| FGM | Fast-Greedy Optimisation of Modularity. |
| FOMD | Fraction Over Median Degree. |
| HPC | Hotspots per Community. |
| LFR | Lancichinetti–Fortunato–Radicchi Benchmark. |
| ODF | Out Degree Fraction. |
| OSN | Online Social Network. |
| RTÉ | Raidió Teilifís Éireann. |
| TPR | Triangle Participation Ratio. |
| UB | Upper Bound. |
| UPC | Users per Hotspot. |

# DECLARATION

I declare that this thesis, titled "*Methods for Defining Dynamic Online Communities and Community Detection in Fast-paced Social Media Streams*", is composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification.

Galway, May 2, 2019



Hugo Fernando Hromic Mayorga

# ACKNOWLEDGEMENTS

First and foremost, I would like to thank my mother *Tatiana* and posthumously my grandparents *Yolanda* and *Svetko*. They are the major reason for who I am today and have always provided me with everything I could ever need in life, including their love, care and values. Specially I would like to thank my mother for never letting me give up on my aspirations and goals in life.

I would also like to thank *Daniela*, my love, best friend and companion in life. You are my favourite person in the whole world, making my life better and happier ever since the first day we got together. I am truly grateful of your patience, love, care, support, advice, points of view, discussions and companionship, all of which were fundamental to complete this thesis.

Special thanks also to my closest friends, *Gabi* and *Erik*, two exceptional people that I have the privilege to be friends with. I have spent countless hours in their company, in happy and not so happy times. Without a doubt, you guys are part of my life and the success of this work.

During my studies in Ireland, I have met a large number of great people that in one way or another, have also contributed to this achievement. In no particular order, I would like to affectionately thank *Marisol*, *Mario*, *Bianca*, *Emir*, *Andrea (B)*, *Andrea (Y)*, *Thu* and *Pablo*, and many others. As you all know, you have been important in the journey of my studies.

I want to thank also my supervisor *Conor*, my external examiner *Miriam* and my internal examiner *John*. All of them contributed to making me a better scientist and professional. Their constant guidance and suggestions were key to the successful completion of this thesis.

I finally want to thank the institute staff for their help and support over the years: *Donal*, *Hilda*, *Claire*, *Michelle*, *Christiane* and *Brian*. You all are the unsung heroes of the institute.

This work was created with the financial aid of Science Foundation Ireland (Grant Number SFI/12/RC/2289), the Clique Strategic Research Cluster (SFI/08/SRC/I1407) and the EU project ROBUST (Grant No. 257859).

# 1 INTRODUCTION

In this chapter, the context and motivation of the thesis are introduced. Afterwards, the problem definition and the research questions that are addressed in this work in relation to the community detection in microblogging are presented. In addition, the thesis contributions are also introduced. Finally, the organisation and overview of the thesis is described.

## 1.1 CONTEXT AND MOTIVATION

During the years, online social networks (OSN) have developed from simple static blogs into richer interactive systems of diverse kind and purpose such as Facebook, Instagram, Twitter, LinkedIn or YouTube, to name a few. One kind of OSN is *microblogging*, which allows for fast, real-time and open broadcasting of short content among friends, acquaintances and followers.

Well known examples of microblogging services are Twitter, Weibo (the Chinese counterpart of Twitter), and Tumblr, which is similar to Twitter but focused on multimedia posts such as photos, videos or audio. Twitter is currently one of the largest microblogging OSN in the world, with more than 330 million monthly active users as of December 2018 [Asl18]. Moreover, an average of 500 million *Tweets* (short messages) per day are generated in the platform.

Compared to other social media, relationships between users in Twitter are also less personal and a more open ambient for socialising is promoted instead [SKC09]. Users in Twitter – and in other microblogging services – can post messages publicly and other users can reply, quote and *retweet* (rebroadcast) them. In addition, users can mention other users in Tweets to gather their attention and a limit of 280 characters per post is enforced, prompting users for brevity and clarity in their content. Moreover, users can choose to follow other users for up-to-date content, often with no approval needed or without requiring to be followed back. However, only 20 % of users tend to follow each other [KLP+10], i.e. a low reciprocity, in contrast to other services such as Flickr (70 % [CMG09]) or Yahoo! 360 (80 % [KNT10]).

While online social networks connect people to friends and acquaintances, there is also an increasing content disorganisation and overload for their users due to the high volume of public messages. Fortunately, this problem can be alleviated with content filtering or clustering techniques such as community detection in networks [AW10; For10; PKV+11; MV13; XKS13]. Using communities, users can be grouped according to common topics, common people or by homophily to help them better focus on what is more relevant for their interests.

However, a fundamental challenge in the community detection literature is the diversity for a definition of *user community*, which makes evaluating and interpreting algorithms difficult. In general, community detection approaches for social networks in the literature propose the following broad hypothesis for user communities in varying degrees.

**DENSITY HYPOTHESIS** ($H_1$) in a social network, a user community corresponds to a distinguishable, locally dense and connected sub-network, that is less connected to the rest.

**NULL HYPOTHESIS** ($H_0$) random networks do not have distinguishable community structures.

In this thesis, the task of community detection and its associated challenges in the context of microblogging online social networks are addressed in both, a static and a dynamic scenario. This type of social media has characteristics that impose challenges for applying standard community detection techniques. Therefore, in this work the structural and temporal characteristics of user communities that can form in microblogging are investigated with the goal of developing improved methods that can better identify communities in this type of social networks.

In addition, real-world prototype applications are introduced that incorporate community detection approaches in the context of microblogging social media – represented by Twitter – and a user activity hotspots model is proposed in this thesis to improve their intended purpose.

## 1.2 COMMUNITY DETECTION IN MICROBLOGGING

One of the most fundamental problems in modern network science is the identification of structural communities in complex networks [For10; FH16; GN02; NG04]. Community detection is a challenging subject because there are many valid interpretation of its aspects, e.g. the definition of community itself, therefore validation and comparison of approaches in various domains is difficult. Understanding communities is important for the study of their functionality in many domains such as social network analysis [MDC+16], biochemistry [GA05] and communication

networks [DFC05]. Finding groups of users with common interests can help not only businesses strategies but also the users themselves to get more connected.

In general, *community detection* or *community discovery* is the task of identifying clusters, modules or communities of users that share something in common. For example, users discussing the same topics or users with similar interests or affiliations. The usual source of information used to discover these communities are complex social networks of users that are connected through observable relationships such as friendships, followers, co-authoring or interactions.

The definition of a user community in Twitter – and therefore for microblogging – is often described in the literature as a group of nodes more densely connected to each other than to nodes outside the group [TL10; JSF+07; GJK12; LB14; SOM10; SKC09; YL15]. However, it is argued in this thesis that, in microblogging social networks, people do not seek to be closely related but instead are more curious about the collective public opinion of the global user-base, unlike other social platforms where strong relationships between users are preferred, e.g. Facebook.

Instead of attempting to craft yet another community definition specifically for microblogging, this thesis will prefer and evaluate a more flexible open and wider interpretation of user communities (Chapter 3). In particular, a differentiation between *functional* and *structural* definitions of user communities is adopted [YL15].

**FUNCTIONAL COMMUNITIES** are defined as groups of users sharing a common and independent social *function*, e.g. fans of the same football team, people living in the same area or discussing the same topic. Social functions such as topic tags, locations or external users referenced in common are independent from any underlying social network, and instead are explicitly stated by the users in their messages.

**STRUCTURAL COMMUNITIES** are defined as groups of users with a particular pattern in their *connectivity in a network*, e.g. their average node degree or clustering coefficient.

It is then argued that functional communities can be uncovered from structural patterns in a network of live interactions. Moreover, functional communities will represent ground-truth information because users themselves explicitly state the social function they use in their posts, e.g. referencing the same hashtag or mentioning the same celebrity.

In this thesis, thirteen commonly used community scoring functions pre-classified into four families are considered for evaluation and real-world Twitter data streams under different settings and periods of time are investigated. Moreover, in this thesis it is proposed that, by identifying user activity *hotspots* in these networks, it is possible to find time periods during which user communities are easier to discover.

Network *sparsity* in this thesis refers to the proportion of connected nodes in a network compared to the total possible number of links. If few nodes are connected, the network is said to be sparse. For example, sparse social networks are these where users are connected in small groups that are highly disconnected from other groups in the network. Furthermore, the concept of *noisy* data refers to undesirable or non-contributing data that has a large amount of additional meaningless information. For example, in a social network, noisy content can consist of posts that link to unwanted web resources or posts that include a high number of unused tags.

Most community detection approaches are designed for static networks [For10; PKV+11]. It is demonstrated in this thesis that standard community detection algorithms are challenged by the fast-paced dynamics and link sparsity of microblogging data (Chapter 4). Therefore, it is proposed to define user activity hotspots that leverage the temporal characteristics of microblogging social networks to improve community detection methods in this medium (Chapter 5). Followers networks, which are comparatively static networks, are commonly used as network data for community detection in microblogging (refer to Section 2.2.2). However, these networks are often prohibitive to capture for global analysis and are not suitable for detecting fast-paced community formations and terminations [DOG15]. Therefore, in this thesis the usage of live streams of interactions is preferred instead.

## 1.3 RESEARCH QUESTIONS

All the research carried in this work is divided into four main stages: (1) a *preliminary* stage, (2) a *static scenario* stage, (3) a *dynamic scenario* stage, and (4) a *practical applications* stage. Each of these research stages propose a main research question and a set of sub-questions which are addressed in the respective chapters of the thesis.

First, in the **preliminary stage**, the construction of ground-truth functional communities and user interactions networks from microblogging social media – represented by Twitter – is studied. The main research question and its sub-questions for this stage are introduced below.

**(RQ1)** How can microblogging ground-truth and structural data for community detection be assembled and modelled in Twitter social streams?

**(RQ1.1)** For constructing ground-truth data, how can independent, explicitly user-labelled, functional communities be modelled in Twitter social streams?

**(RQ1.2)** For constructing structural data associated to the ground-truth in (RQ1.1), how can networks of user interactions be modelled in Twitter social streams?

**(RQ1.3)** What are the global properties, e.g. size and membership distributions, of the defined ground-truth functional communities in (RQ1.1) and (RQ1.2)?

These are addressed in Chapter 3, where the proposed ground-truth functional communities can emerge from live microblogging streams of user interactions such as posting, replying, user mentioning, posts rebroadcasting and content quoting. In addition, these user interactions can be also used to build a network of users interacting with each other at different points in time.

Afterwards, in the **static scenario stage**, the constructed ground-truth functional communities and interactions networks are characterised and evaluated using a number of structural scoring functions for communities in a static context, i.e. where these communities are considered as a single whole unit, regardless of their user activity in time. The main research question and its sub-questions for this stage are introduced below.

**(RQ2)** How do existing structural community definitions accommodate to microblogging ground-truth communities, including their robustness to random perturbations?

**(RQ2.1)** Do the defined ground-truth functional communities in (RQ1.1) evidence distinctive characteristics of structural communities, i.e. higher clustering coefficient, average degree, edge density and cohesiveness, in the associated networks of user interactions in (RQ1.2) in comparison to random groups with similar size and shortest-path distribution?

**(RQ2.2)** How well do state-of-the-art structural community definitions, e.g. based on triangle participation, conductance or modularity, align to the defined ground-truth functional communities in (RQ1.1) and (RQ1.2), including their robustness to random perturbations, e.g. member swapping, shrinking or expansion?

In this static scenario stage, which is investigated in detail in Chapter 4, multiple state-of-the-art structural scoring functions for communities commonly used in the literature are evaluated over the defined functional ground-truth in (RQ1). For instance, the fraction over median degree, triangle participation ratio, cut ratio, conductance, maximum out-degree fraction and modularity. Moreover, also based on their structure, the ground-truth communities are evaluated in terms of defined community goodness metrics: clustering coefficient, edge density, cohesiveness and separability. The noisy environment of microblogging is challenging for structural definitions, therefore multiple perturbation strategies for communities, i.e. node swapping, random replacement, group expansion and group shrinking are also evaluated.

The main motivation of this thesis lies in that it is difficult to identify user communities in live streams of microblogging social interactions due to their velocity and low density characteristics. Moreover, it is investigated that the highly dynamic and fast-paced nature of microblogging causes users to switch topics or lose interest quickly, rendering standard community discovery approaches based on more static and dense networks less effective.

Therefore, a third **dynamic scenario stage** is also proposed, where the temporal dynamics of the ground-truth functional communities based on the interactions networks are characterised, evaluated and contrasted to the static scenario in (RQ2). The main research question and its sub-questions for this stage are introduced below and investigated in detail in Chapter 5.

**(RQ3)** How can activity hotspots based on the dynamic user activity in time be identified in the defined ground-truth communities to improve community detection?

**(RQ3.1)** What are the temporal characteristics, for instance the user activity distributions, of the defined ground-truth functional communities in (RQ1.1) and (RQ1.2)?

**(RQ3.2)** Using the dynamic user activity in time as a basis, how can activity *hotspots* be identified in the defined ground-truth functional communities in (RQ1.1) and (RQ1.2) to be used for further identifying time-scoped sub-communities?

**(RQ3.3)** Considering the identified time-scoped sub-communities based on user activity hotspots defined in (RQ3.2), how well do the state-of-the-art structural community definitions investigated in (RQ2.2) now align to these sub-communities in comparison to the ground-truth functional communities in the static scenario, i.e. without considering their user activity context?

It is observed during (RQ3.1) that the defined ground-truth functional communities in (RQ1) can be highly dynamic in time for microblogging. If these temporal characteristics are not taken into account, the functional communities become structurally intricate and more difficult to distinguish in long periods of time. Therefore, it is necessary to identify the particular periods in time of the communities in which their user activity is at its prime. These moments in the life-cycle of the communities are then considered as user activity *hotspots*, that once identified, improve the community goodness metrics individually in contrast to their parent community.

Finally, the last stage of the research is the **practical applications stage**. In this stage, the applicability of all the findings in the previous stages is discussed. The proposed main research question for this stage, explored in detail in Chapter 6, is introduced below.

**(RQ4)** How can the findings from (RQ3) be integrated into real-world practical applications designed for different types of microblogging users, that utilise common community detection methods over microblogging data in their workflow?

The real-world practical applications that motivate this work are divided in two classes: (1) applications for *end-users*, aimed and designed for users of the microblogging social platform itself, and (2) applications for *decision makers*, focused instead in supporting enterprise community owners and managers. For each of these application classes, a set of motivating prototype applications developed for this thesis are introduced and discussed in Chapter 6. Furthermore, the applicability of the proposed model of temporal dynamics for microblogging user communities and how it can improve the intended purpose of these applications is discussed.

From all the above research questions and sub-questions, a set of fundamental research outcomes are expected from the investigation work in this thesis, as described below.

- Ground-truth communities defined using social functions have distinctive structural patterns in a network of live user interactions in microblogging. From: (RQ2.1).

- Microblogging users switch topics quickly and do not participate in steady and long-lived communities, rendering conventional structural community discovery approaches based on static and dense networks less effective in microblogging. From: (RQ1.3) and (RQ2.2).

- Activity hotspots in underlying user interactions networks from microblogging can be identified for ground-truth communities defined using social functions. From these hotspots, time-scoped functional sub-communities can be considered. Then, structural community definitions better align to these temporal functional sub-communities individually than to the whole original non-separated communities. From: (RQ3.1), (RQ3.2) and (RQ3.3).

Evidence to support all of the above outcomes is provided in each relevant chapter of this work. Therefore a better understanding of how functional user communities in microblogging can be uncovered from the structure of live streams of user interactions that may originate them. Furthermore, improving the task of community detection for microblogging by leveraging its temporal dynamics can also provide a range of useful and interesting applications, which can be applied to this social media for the benefit of both, end-users and decision makers alike.

## 1.4 THESIS CONTRIBUTIONS

Addressing the research questions of this work provides a number of contributions to the research field of community detection for microblogging, including dynamic community detection, visualisation and evaluation. The identified contributions of this thesis are below.

1. From (RQ1), a methodology for building ground-truth functional communities from microblogging live user interactions. In particular for stream-based Twitter datasets.
2. From (RQ2) and (RQ3), an in-depth characterisation, understanding and evaluation of global and structural properties for functional communities in microblogging social media, for both the static and dynamic scenarios.
3. From (RQ2) and (RQ3), a set of recommendations on community detection algorithms based on data-driven evaluation of Twitter user interactions networks.
4. From (RQ3.2), a strategy for the identification of temporal activity hotspots in functional communities in microblogging based on the network of user interactions, that improves the performance of existing community detection algorithms designed for the static data.
5. From (RQ4), a set of three motivating demonstration applications – two for end-users and one for decision makers – designed for microblogging social media, including a community detection system and a dynamic communities visualisation tool.
6. An open source implementation of the analytical framework developed for this thesis[1].

### 1.4.1 Publications

Furthermore, the following publications emanated during the development of this thesis.

- H. Hromic and C. Hayes. "Characterising and Evaluating Online Communities from Live Microblogging User Interactions". In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. Aug. 2018, pp. 21–24
- H. Hromic and C. Hayes. "Visualising the Evolution of Dynamic Communities in Social Networks using Timelines". In: *3nd ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data (AALTD)*. The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD). Sept. 2018

1 Online public repository: https://github.com/hhromic/dyncom-analytics

- H. Hromic, A. Barraza-Urbina, C. Hayes, et al. "Mining TV Twitter Networks for Adaptive Content Navigation and Community Awareness". In: *Expert Update*. AI-2016 Thirty-sixth SGAI International Conference on Artificial Intelligence 17.1 (2017)
- H. Hromic, N. Prangnawarat, I. Hulpuş, et al. "Graph-Based Methods for Clustering Topics of Interest in Twitter". In: *Engineering the Web in the Big Data Era*. Ed. by P. Cimiano, F. Frasincar, G.-J. Houben, et al. Lecture Notes in Computer Science. Springer International Publishing, 2015, pp. 701–704. ISBN: 978-3-319-19890-3
- H. Hromic and C. Hayes. "Constructing Twitter Datasets using Signals for Event Detection Evaluation". In: *Synergies of Case-Based Reasoning and Data Mining Workshop*. 22nd International Conference on Case-Based Reasoning. Sept. 29, 2014
- H. Hromic, M. Karnstedt, M. Wang, et al. "Event Panning in a Stream of Big Data". In: *LWA Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML)*. 2012

## 1.5 THESIS OVERVIEW

This thesis is organised as follows. First, the background and the position of the thesis in the current literature are established in Chapter 2. Then, the proposed approach for building ground-truth communities and interaction networks from Twitter is described in Chapter 3.

Afterwards, a data-driven characterisation of microblogging communities in a static scenario is in Chapter 4. The investigation on temporal community detection is in Chapter 5, alongside a characterisation of the improved ground-truth communities. The real-world practical applications of the proposed temporal method are in Chapter 6, and the overall conclusions and future directions of the thesis are presented in Chapter 7.

# 2 BACKGROUND AND RELATED WORK

In this chapter, the foundations and position of the thesis are discussed. First, the fundamental concepts used in the thesis are introduced and explored, then a systematic review of the recent literature involving the problem of community detection in general and in the context of microblogging social media is presented. Afterwards, the characteristics and social functionalities of microblogging that motivate the interest of investigating community detection for this particular medium are discussed. Lastly, the research methodology of the thesis is presented.

The research questions, sub-questions and outcomes of this thesis – introduced in Chapter 1 – focus on the investigation of the problem of community detection, the characterisation of user communities, the evaluation of different detection approaches and the improvement of existing techniques in the context of microblogging social networks.

## 2.1 THESIS FOUNDATIONS

The research carried in this thesis is supported on three foundations: (1) **social networks**, (2) **user communities** in these social networks, and (3) the **detection** of these user communities in the social networks. All of these topics are considered in the context of microblogging social media and are the basis for the research questions proposed in Chapter 1.

### 2.1.1 Social Networks

Social networks are the fundamental organisational structures on which the research of this thesis are supported. Investigating social networks to identify local and global patterns, influential entities, and explore network dynamics is an interdisciplinary academic field that emerged from sociology, social psychology, statistics, and graph theory.

**SOCIAL NETWORK** [Was94]

> A *social network* organises social actors such as individuals or organisations according to their relationships and interactions, enabling for a set of methods able to analyse the structure of whole social entities and the application of an array of theories explaining the patterns observed in these structures.

Early structural theories in sociology that emphasise the dynamics of dyads and triads of social actors, including group affiliations were proposed by Georg Simmel – a German sociologist and philosopher – in 1922 [SD03]. Shortly after in the 1930s, Jacob Moreno – a leading Romanian-American psychiatrist, psycho-sociologist and social scientist – is credited with developing the first instruments to study interpersonal relationships in social networks.

These early ideas were later formalised in the 1950s. Furthermore, the theories and methods of social networks became more widespread in the social and behavioural sciences by the 1980s [Was94; Fre04]. Social network analysis is currently one of the major fields in contemporary sociology, being also adopted in other social and formal sciences. In the late 1990s, social network analysis was further advanced by work from sociologists, political scientists, and physicists such as Duncan J. Watts [WS98] and Albert-László Barabási [BA99]. Social networks and other complex networks form part of the field of network science [BMB+09; EK10].

In the context of the Internet and online services, social networks are also very relevant because they can be used to model personal relations of internet users. The term **social graph** was popularised on May 24, 2007 at the Facebook F8 conference[1] where the newly introduced Facebook Platform – a major player in the social media landscape – was set to utilise the user-provided relationships between Facebook users to offer them a better online experience. These social graphs have been further expanded to cover not only Facebook users but also users in other social media. In this online setting, social graphs are then the formal representation of these social networks. In this manner, the notion of *graph* from graph theory enables a number of graph-based approaches to operate on social networks as well.

Social networks are in general considered to be self-organising, emergent and complex. Global patterns can appear from the interaction of the nodes in the network and become more apparent when the network grows [Wel08; NBW11]. Nevertheless, in the extreme case of considering all interpersonal relationships in the world, a global network analysis then becomes unfeasible and would be so overloaded that it would be rather uninformative [WF94]. In addition, practical limitations in computing resources, ethical concerns and participant recruitment/rewarding also prevent social network analysis at such a global scale [Kad12; Gra76].

1 https://newsroom.fb.com/news/2007/05/facebook-unveils-platform-for-developers-of-social-applications/

Potential local patterns that may exist within the global network may be lost in a large network analysis, therefore the information quality of the network may be more important than its scale for understanding its properties. For this reason, social networks are investigated at a scale relevant to the research questions under study. In general, there are three abstractions for performing social network analyses: the (1) **micro**, (2) **meso**, and (3) **macro** levels.

#### MICRO LEVEL ANALYSIS

Research at the micro level of social networks typically begins with a single individual and expanding when her social relationships are further explored. Also it can begin using a small group of individuals for a particular social context. Analyses at the micro level are often carried using more fine-grainedd sub-levels: the (1) *dyadic*, i.e. only between two individuals, (2) *triadic*, i.e. between three individuals or more [Kad12; KCH+13], (3) *actor*, e.g. using egonetworks [JV11; ML14], and (4) *mixed*, i.e. beginning at the micro level but may progress to the meso level [Noo14].

#### MESO LEVEL ANALYSIS

Research at the meso level of social networks begins with a population size between the micro up to macro levels. Moreover, meso level can also refer to research designed specifically to discover connections between micro and macro levels. Meso level networks are of low density and causal processes not at the interpersonal level may exist [HSSoo]. Analyses at the meso level are often carried on the following types of social networks: (1) *organisations*, i.e. social groups with a collective goal [RNo7; SC17], (2) *randomly distributed networks*, i.e. those that model common attributes of human social networks such as reciprocity, homophily[2] and transitivity [MSCo1; CD11], and (3) *scale-free networks*, i.e. those with a degree distribution that follows a power law and therefore containing few central individuals with node degrees above the average [Baro3; MPC+o6; CHD10].

#### MACRO LEVEL ANALYSIS

Research at the macro level of social networks study the global outcomes of the individuals rather than their interpersonal interactions, e.g. the economic and political effects over a very large population. Analyses at the macro level are often carried on the following types of social networks: (1) *large-scale networks*, i.e. those that model global-scale phenomena such as the stock market, and (2) *complex networks*, i.e. those that model substantial non-trivial topological features such as heavy tail in the degree distribution, high clustering coefficient or community structures [Stro1; BLH12]..

2 Homophily is the tendency of individuals to associate with similar others, *birds of a feather flock together*

In this thesis, the proposed research questions and sub-questions contemplate the analysis of microblogging social networks at the meso and macro level, where these networks are investigated considering them as randomly distributed, scale-free and complex.

Finally, when computer social networking software is combined with global-scale computer networks such as the Internet, a new medium for social interaction emerges [AH17]: *social media*. Relationships between individuals in social media can be characterised by context, direction and strength, and the content of these relations is very often a resource that is exchanged by the participants of the network. In the context of online communication, these resources can be abstracted in many ways, e.g. a data file, a computer program or in a more social context, the provision of emotional support or the arrangement of a meeting. Moreover, with the introduction of electronic commerce, these resources were extended to money transactions, goods or services [GHW97]. In general, all these resources can be considered as *social objects* that connect people together to truly become social, otherwise they loose interest to remain in contact [Eng05].

Social network analysis methods have become essential to investigate online social media. Furthermore, the large size and fast-paced nature of social media has motivated new network metrics. However, a key challenge in social networks analysis for social media are the unreliable observations of these network metrics in missing, sparse and noisy data [WJL+16].

In this thesis, **two** research sub-questions specifically involving social networks in microblogging social media are investigated in Chapter 3 and Chapter 4.

1. (RQ1.2) → For constructing structural data associated to the ground-truth in (RQ1.1), how can networks of user interactions be modelled in Twitter social streams?

2. (RQ2.1) → Do the defined ground-truth functional communities in (RQ1.1) evidence distinctive characteristics of structural communities, i.e. higher clustering coefficient, average degree, edge density and cohesiveness, in the associated networks of user interactions in (RQ1.2) in comparison to random groups with similar size and shortest-path distribution?

### 2.1.2 User Communities

Since the early days of social network research, the study of individuals organised in communities has been of interest [Bar54]. In this seminal work, a community was considered in a specific geographic location where the ties between people were investigated. In particular, who talked, associated, traded and attended church with whom was studied.

In modern social networks, however, telecommunication devices and online social media services have enabled the formation of *virtual* communities of individuals and organisations. Due to the global scale of such services, large and more complex analysis and maintenance techniques are necessary, often requiring network science approaches. Therefore, advanced community development research nowadays also makes extensive use of such methods.

The formation of real and virtual communities of individuals can be explained using the concept of *homophily*, which is the tendency of individuals to associate with similar others, e.g. *"birds of a feather flock together"* [LM54; MSC01]. Evidence of homophily in different forms has been reported, where it is established that similarity such as age [Mar88], gender, class and organisational roles facilitates connection among individuals [MSC01]. Furthermore, other common characteristics identified for homophilic similarity are beliefs, values and educational background. Research also have suggested that homophily among individuals helps them on better accessing information resources due to collaboration on the same interests [DSJ+10].

In the context of online social media, homophily is also present. Online social services are highly competitive and often cultivate an homophilic ambience to retain their users. For example, in Facebook, when users manifest interest or interact with an article of a certain subject, the service will continue to present similar resources to them under the assumption that they will also be interested on this additional content. Homophily can also isolate users of social media services because individuals with similar ideologies tend to only interact with each other.

#### COMMUNITY

In general terms, a *community* is a small or large group of individuals that have some aspect in common, such as norms, religion, values, or identity. The individuals in these groups often share a *sense of community* that brings them together [MC86]. For example, their physical location such as a given geographical place (e.g. country, village, town, or neighbourhood), or a virtual space through online communication services. Moreover, people often value the sense of community for their own identity and their roles in social institutions, e.g. family, work, government or society in general [JNH+12]. Communities can be small, relative to the personal social ties of the individuals, or large, relative to bigger national communities, international communities and online virtual communities [Jam06].

#### ONLINE COMMUNITY

One special type of communities are *online virtual communities*, which have particular characteristics because of the communication channels that support them, e.g. social media services or modern telecommunication systems. Online communities enable new dynam-

ics compared to conventional real-world communities, which are more limited in this context. For instance, virtual communities allow for user anonymity, which convey weaker bonds than actual real-world communities. Conversely, social media that offer online interaction between individuals such as Twitter, Facebook, Instagram or Tumblr, easily allow more undesirable behaviour, e.g. users stalking and bullying of others in their community. Individuals in online virtual communities should be considered more as an anonymous audience with fewer ties among them than a big, cordial, traditional community. Likewise, traditional real-world communities should not be considered as a small audience because they have social density properties that audiences lack [Shi09].

#### COMMUNITY STRUCTURE

In the context of social networks – or social graphs –, a network $G = (V, E)$ is said to contain *community structures* if the nodes $v \in V$ of the network $G$ can be easily grouped into sets of nodes $C_i \cup V$ such that each set of nodes $C_i$ is more densely connected internally in $E$ than with the rest of the network [GN02; For10; MV13; POM09; FB17]. Furthermore, these sets $C_i$ can be overlapping or non-overlapping [XKS13; PDF+05]. The latter is a useful simplification in most studies because most community detection approaches are designed for this type of community structure. However, a better representation is sometimes necessary and overlapping nodes are permitted. For example, in a social network where each node is an individual and communities represent different groups of people, it is possible to have one community for family, another for co-workers or one for friends from school [Bra02]. Community structures are based on the idea that pairs of nodes are more likely to be connected if they are both members of the same community or communities, and less likely to be connected if they do not share any communities [RCC+04], similar to homophily.

Community structures allow to create large-scale maps for complex networks. Individual communities can be interpreted as node groupings in the networks that make their study easier [New06b]. Furthermore, individual communities also can help on the identification of the functions of the system modelled by the underlying network because community structures often correspond to the actual functional elements of the system under study [YL15]. For example, in metabolic biological networks, such functional units correspond to cycles or pathways and in protein-protein interaction networks, community structures correspond to proteins with similar functions inside a biological cell [AEC+15]. In the same manner, engineering networks can form community structures by design attributes [Bra02] and citation networks can form communities

by research topic [BLH12; GN02]. Therefore, the discovery of these structures can provide important insights about the relationship between the network functions and the network topology.

Community structures are important in network science because they often exhibit different properties when compared to the average properties of the complex networks under study. If only the average properties were to be considered for analysis, the rather important and interesting features inside these networks would be discarded. For instance, in the context of social networks, both gregarious and reticent groups might exist at the same time [New06b]. However, not all complex networks have meaningful community structures. For example, the Erdős–Rényi (based on random graphs) [ER59] and the Barabási–Albert model (based on preferential attachment) [BA99], do not evidence community structures.

In this thesis, **two** research sub-questions specifically involving community structures in microblogging social networks are investigated in Chapter 3 and Chapter 4.

1. (RQ1.1) → For constructing ground-truth data, how can independent, explicitly user-labelled, functional communities be modelled in Twitter social streams?

2. (RQ2.1) → Do the defined ground-truth functional communities in (RQ1.1) evidence distinctive characteristics of structural communities, i.e. higher clustering coefficient, average degree, edge density and cohesiveness, in the associated networks of user interactions in (RQ1.2) in comparison to random groups with similar size and shortest-path distribution?

### 2.1.3 Community Detection

One of the most fundamental problems in modern network science is the identification of structural communities in complex networks [For10; FH16; GN02; NG04]. This problem is often referred to as *community detection* or *community discovery*. User communities are traditionally seen as groups of vertices in a network that have higher chances of being connected to each other than to other vertices, however other definitions are also possible. Community detection is a challenging subject because there are many valid interpretation of its aspects, e.g. the definition of community itself, therefore validation and comparison of approaches in various domains is difficult. Understanding communities is important for the study of their functionality in many domains such as social network analysis [MDC+16], biochemistry [GA05] and communication networks [DFC05]. Finding groups of users with common interests can help not only businesses strategies but also the users themselves to get more connected.

For the case of social networks, two types of user groups can be defined: (1) *explicit groups*, i.e. formed by user subscriptions, and (2) *implicit groups*, i.e. implicitly formed by user social interactions. In community detection, the goal is to discover implicit groups in a complex network, i.e. the memberships of the participating users are not explicitly given. The network can be of any nature, i.e. biological, engineering or citation networks, however is common to apply community detection to social networks (refer to Section 2.1.1). The definition of user community for this task can be subjective and very diverse (refer to Section 2.1.2). Moreover, in recent years community detection has been strongly focused on *online* user communities.

Community detection methods for complex networks can be classified into four major families, based on how they define the communities under study and the type of connectivity they consider: (1) only internal connectivity, (2) only external connectivity, (3) both, internal and external connectivity, and (4) based on network models. In this thesis, a collection of thirteen different structural community measures, i.e. community definitions, belonging to all of the above families are investigated for the case of microblogging social networks.

#### COMMUNITY DETECTION IN SOCIAL NETWORKS

> The task of *community detection in social networks* is defined as finding implicit groups of individuals $u \in V$ in a social network $G = (V, E)$, implicitly formed by their social interactions $e \in E$. In this context, the candidate groups can be measured using a structural scoring function $f(C)$ that quantifies the structural qualities of interest for the analyst.

Some social media services, e.g. Facebook, allow for their users to join and manage virtual groups, therefore extracting implicit groups based on their interactions is less necessary. Moreover, not all individuals are willing to make efforts to join available groups. However, not all social media provide an explicit community-aware platform, e.g. Twitter, and groups can change dynamically. In these cases, community detection is a valuable analytical tool.

An important open question in the community detection literature is the evaluation of methods and algorithms for this task. Evaluation methodologies must consider networks with known structures [MV13; HBG+14; LC14; YL15], and the most common approach is to generate synthetic random networks with these patterns. For example, a random network can be generated containing four equally-sized partitions and the probabilities of the connections inside and between these partitions can be adjusted to create community structures of varying complexity for the detection algorithm under evaluation [RCC+04]. Such artificial random networks are known as *planted l–partition models* [CK01] or, more generally, as *stochastic block models*. All these network models are designed to contain patterns resembling community structures.

Other evaluation approaches have been proposed as well that are more flexible. For example, the popular Lancichinetti–Fortunato–Radicchi (LFR) benchmark allows for nontrivial degree distributions and varying community sizes [LFR08], in contrast to the simpler four partitions method, allowing to evaluate community detection methods under more difficult scenarios. Another example is the Pasta and Zaidi method that evaluates community detection approaches using evolution dynamics instead of static network configuration models [PZ17].

The typical process of evaluating community detection using synthetic networks begins with a network containing well-defined structural communities. Afterwards, these structures are degraded gradually by re-arranging or removing links between nodes, thus detecting these structures becomes increasingly more difficult. The structures found by the algorithm under study are compared to the community structures in the generated networks. The performance of the community detection approach under evaluation can be measured using metrics from information theory such as normalized mutual information or variation of information [CT12].

Ground-truth user communities in this thesis are defined through explicitly labelled social functions in microblogging social media such as topic tags, locations or external users referenced in common. Furthermore, community detection methods are defined through structural scoring functions that measure different structural aspects of communities such as conductance, clique formation or modularity. Community detection evaluation is then defined as quantifying the ability of these structural definitions to model the functional ground-truth communities.

In this thesis, **two** research sub-questions specifically involving the evaluation of community detection approaches represented as structural scores in the context of microblogging social networks are investigated in Chapter 4 and Chapter 5.

1. (RQ2.2) → How well do state-of-the-art structural community definitions, e.g. based on triangle participation, conductance or modularity, align to the defined ground-truth functional communities in (RQ1.1) and (RQ1.2), including their robustness to random perturbations, e.g. member swapping, shrinking or expansion? (static scenario)

2. (RQ3.3) → Considering the identified time-scoped sub-communities based on user activity hotspots defined in (RQ3.2), how well do the state-of-the-art structural community definitions investigated in (RQ2.2) now align to these sub-communities in comparison to the ground-truth functional communities in the static scenario, i.e. without considering their user activity context? (dynamic scenario)

## 2.2 COMMUNITY DETECTION IN SOCIAL NETWORKS

In social networks, users develop natural groupings from finding other users with similar interests. This phenomenon is known as homophily (refer to Section 2.1.2) and these groupings are often called modules or communities (refer to Section 2.1.3. Finding community structures within complex networks can be a computationally difficult task. The number of communities present in the complex network under study is likely unknown and the community structures are often of uneven size and density. Nonetheless, several approaches for community detection have been developed and applied with varying levels of success in the literature [POM09].

In this section, a systematic review of the recent literature involving the problem of community detection in general and in the context of microblogging social networks is presented. This thesis addresses both, the problem of defining user communities and the problem of evaluating community detection in the context of microblogging social networks. Furthermore, the dynamic temporal aspects of user interactions are also incorporated by this thesis into the community detection task to further improve the performance and understanding of fast-paced user communities that can form in microblogging social media.

### 2.2.1 General Community Detection

The goal of community detection methods is to identify community structures in a complex network of interest, where nodes represent individuals and edges represent the interaction or similarity between them (refer to Section2.1.3). Community structures are also called modules or clusters and are a subset of nodes in the complex network under study, and often are defined as nodes whose links are denser compared to the rest of the network [NG04].

Community detection approaches based on the analysis of cliques, node degree, and matrix-perturbation have been proposed for identifying cohesive structures from social networks [WF94; FH16; CGP11; MV13; PKV+11; PRS11; XKS13]. Detected structures can range from communities of scientists working on similar areas of research [GN02], i.e. citation networks, biological networks [AEC+15], to individuals who have some common interests [FH16].

The task of community detection can be also considered to be closely related to the task of data clustering. Community detection can be defined in terms of a distance function and a clustering objective function to generate a clustering assignment for each node in the graph to a set of clusters, modules or communities. Nonetheless, while there are similarities between

community detection and data clustering, the former focuses more on the relationship between the nodes in the network, or more generally on the topology of the network.

Research on community detection includes measures for quantifying community structures in networks, e.g. clustering coefficient [WS98; New03] and cohesiveness [LLM10]. A variety of methods have been further proposed: modularity-based methods [BGL+08; NG04], graph cut-based methods [FLG00], spectral clustering [CG97], graph Laplacian-based methods [New06b] and information-theoretic models [RB07]. The dynamic properties of communities in complex networks have also been investigated in the literature [LLD+08; POM09].

One important challenge in community detection is the evaluation of detection approaches and algorithms (refer to Section 2.1.3). For a fair comparison, the definitions of community under evaluation and their interpretations must be comparable across approaches. Moreover, ground-truth data is not easily obtainable, specially for large-scale evaluations. In some cases, it is possible to obtain communities that satisfy cluster validation criteria, however these might not necessarily be supported by real human interactions. In this thesis, ground-truth community data is assembled from explicitly-labelled social functions in real-world microblogging streams, independently of any community detection method, ensuring their validity for evaluation.

For social networks, a number of works have been published in the field of community detection and its applications. These works can be organised by the subject of community detection that they focus for investigation: (1) structural properties, (2) community properties, (3) evaluation methods, and (4) community dynamics.

Works investigating the structural properties of communities that can be found in complex social networks has been proposed [GN02; NG04; GA05; PZ17]. Girvan and M. E. J. Newman propose an approach for detecting communities that uses node centrality indices to find the community boundaries [GN02]. They use both, synthetic and annotated real-world networks to evaluate their method and demonstrate that it is able identify communities reliably. Furthermore, they also investigate non-annotated real-world networks and meaningful network partitions were still successfully discovered. Later, M. Newman further investigate using betweenness centrality, i.e. nodes that act as bridges, for community detection [New03]. This approach iteratively removes edges from the network based on betweenness measures to form community structures within the network under study. After each removal iteration, the node betweenness metrics are recomputed. In addition, the authors provide a measure of the quality of the discovered community structures. This measure is then used to select the optimum number of communities into which divide the input social network. These methods are also evaluated in both, synthetic and real-world data.

Guimerà and Amaral recognise the importance of aligning structural communities to their underlying social functions [GA05]. It is demonstrated that modules, e.g. communities, can be found in a number of real-world complex networks and further classify the individuals in these communities according to their roles in the structure, computed from structural patterns inside and outside the found communities. Their approach is presented as a *cartographic representation* of complex networks. Pasta and Zaidi delve further in the topological characteristics of social networks for community detection [PZ17]. Social networks having different topologies are investigated by evaluating community detection algorithms using specially generated synthetic networks. They adopt the Lancichinetti-Fortunato-Radicchi (LFR) model and extend it into a new model named *Naïve Scale-Free Clustering*, that actively counters for potential bias produced by the LFR generation model. Their results suggest that current popular community finding algorithms are limited when network topologies are too different. The authors highlight the need to further investigate current algorithms to improve their performance. This thesis gives particular attention to the social functions that define user communities in microblogging social networks by investigating their alignment to their underlying user interaction topologies.

Further works investigating structural community properties have been also proposed [CK01; RCC+04; PDF+05; RB08; For10; HBG+14]. Condon and Karp studied the problem of partitioning undirected complex networks into two partitions of equal size while minimising the number of crossing edges between them [CK01]. The authors introduce an efficient, i.e. linear time, algorithm designed for this problem and describe it using formal methods in random networks. Their model partitions the nodes of a network into $l$ groups, each of size $n/l$ nodes and $n$ the total number of nodes in the network. Then, probabilities are assigned to each pair of nodes for both, their linking and disconnection. It is then demonstrated that their algorithm is able to find optimal partitions with a known probability and error constant. Radicchi, Castellano, Cecconi, et al. later recognised the lack of quantitative community definitions in most community detection algorithms of the time, making interpretation of their results difficult in the absence of non-network information [RCC+04]. These concerns are addressed by showing how quantitative community definitions can be implemented for existing detection approaches, while enabling self-containment for these methods. Furthermore, a community detection algorithm is proposed with less computational requirements and retaining the same reliability as other algorithms.

Palla, Derényi, Farkas, et al. introduced in 2005 one of the most popular detection methods for overlapping communities in the literature: the *Clique Percolation Method* (CPM) [PDF+05]. This method incrementally construct communities from $k-$cliques, i.e. fully connected sub-graphs of $k$ nodes. In this approach, any two given $k-$cliques are considered adjacent if they share

the same $k-1$ nodes and the maximal union of $k-$cliques that can be reached from each other through a series of adjacent $k-$cliques is defined as a module or community. Overlapping communities can be naturally modelled using this approach and furthermore the authors report that these overlaps can be statistically significant. Their experiments also reveal universal features of real-world networks such as collaboration, word-association and protein interaction networks. Another widely popular model in the literature was introduced by Rosvall and Bergstrom in 2008: InfoMap [RB08]. This algorithm is derived from information theoretic approaches and can be applied to weighted and directed social networks. The probability flow of random walks in a random network is computed and correlated to the flow of information in the network under study. The network is then decomposed into modules by compressing a statistical description of the computed probability flow. The final result is a mapping that simplifies and highlights the regularities in the structure of the network, including their relationships.

Fortunato in 2010 published a fundamental review regarding community detection and its evaluation in complex networks [For10]. The author evaluates eight novel and five traditional community detection algorithms for overlapping and non-overlapping communities on large-scale, synthetic and real-world complex networks with known ground-truth communities. All the studied algorithms were empirically compared using goodness metrics that measure the structural properties of the identified communities, as well as performance metrics that evaluate these communities against the known ground-truth. The reported results show that these two types of metrics are not equivalent. An algorithm may perform well in terms of goodness metrics, but unsatisfactory in terms of performance metrics. The opposite was also demonstrated to hold true. Harenberg, Bello, Gjeltema, et al. in 2014 further reviewed community detection in complex networks, with a focus on statistical methods such as the significance of detected communities [HBG+14]. Guidelines for how methods should be evaluated and compared against each other and applications to real networks were also discussed.

The problem of evaluating methods for community detection in complex networks have been extensively studied in recent years [LFR08; MV13; LC14; YL15; FH16]. For instance, Lancichinetti, Fortunato, and Radicchi proposed in 2008 one of the most recognised synthetic benchmarks: the Lancichinetti-Fortunato-Radicchi (LFR) benchmarking models [LFR08]. Their models account for the heterogeneity of node degree distributions and community sizes. As a demonstration, the state of the art modularity optimisation technique was evaluated using the proposed generative models. The LFR benchmark is a thorough test for algorithms than other standard evaluations, able to reveal limits in community detection approaches that may not be apparent in preliminary analyses. Malliaros and Vazirgiannis further present methods and metrics for

evaluating community detection results in the context of directed complex networks [MV13]. Two classes of metrics are proposed, one focused on the methodological principles of the detection algorithms under evaluation, and another focused on the properties of a good community in a directed network. Lee and Cunningham argue that evaluation based on small and large networks are so different that benchmarking algorithms only on smaller networks provide very little significance on how they would perform on larger datasets [LC14]. Furthermore, annotated ground-truth communities are limited for two reasons: (1) they may only be a small fraction of the real communities in the network, and (2) they not necessarily must correspond to realistic communities in the same network. In this thesis, ground-truth communities for evaluation are generated from large-scale streams of microblogging data (refer to Chapter 3), and the annotations – based on social functions – are provided by users in real-world activities.

Further on the subject of evaluation, J. Yang and Leskovec argue that the goal of network community detection can be thought as the extraction of functional communities based on the connectivity structure of the nodes in the network [YL15]. The authors identify real-world networks with explicitly labelled functional communities that are considered as community ground-truth data. A systematic methodology is then proposed to quantitatively evaluate community detection algorithms based on their definitions of structural community and their alignment to the defined functional ground-truth. Furthermore, they propose a method for detecting communities from a single seed node that extends the local spectral clustering algorithm [CG97]. This thesis borrows a number of conceptualisations proposed in this work (refer to Section 2.4). Lastly, Fortunato and Hric in 2016 further updated their previous work [For10] on community detection approaches and their evaluation [FH16]. In this work, new validation strategies (artificial benchmarks, partition similarity measures, detectability and community structures in real networks) and community detection methods (based on statistical inference, optimisation and dynamics) were incorporated and discussed. The authors highlight that the problem of community detection can be defined in many different and valid ways, therefore there is no definitive answer to which methods are best for every case, despite the fact that many approaches are based on similar ideas. Instead, it is suggested that each community detection scenario should be particularly characterised and an appropriate evaluation and detection methodology selected accordingly.

Processing large social networks for community detection often requires considering the community dynamics in time [DFC05; POM09; MDC+16]. Diesner, Frantz, and Carley studied the Enron email corpus – which contains rich temporal data of internal communication within a large, real-world corporation facing a severe accountancy scandal crisis – using a community

perspective [DFC05]. The authors enhanced the original email database and explored the dynamics of the structure and properties of the communications network. Their findings provide valuable insight into the dynamics of complex social networks, which can be further used for validating or developing community detection approaches. Furthermore, Porter, Onnela, and Mucha highlight the connections of dynamic community detection to problems in statistical physics and computational optimisation [POM09]. Lastly, Musciotto, Delpriori, Castagno, et al. investigate highly sparse temporal networks that model social interactions such as the physical proximity of participants in conferences [MDC+16]. Temporal networks can become very sparse and the data is of lower quality when users restrict their interactions due to privacy concerns. Nevertheless, it is demonstrates that significant information can be still discovered from dynamic temporal networks such as the correlation of events happening during a conference and stable communities interacting over time. In this thesis, different types of functional communities from microblogging social networks are constructed (refer to Section 3). Three of these types are location-based functions and they also are affected by higher sparsity and lower data quality (refer to Chapter 4). However, community structures are still possible to be discovered.

A summary of the reviewed literature for community detection in general complex networks can be seen in Table 2.1 In the table, the community detection approach types, methods and network data they are evaluated on are observed.

### 2.2.2 Community Detection in Microblogging

The definition of a user community for Twitter – and by extension for microblogging – is generally described in the literature as a group of nodes more densely connected to each other than to nodes outside the group [TL10; JSF+07; GJK12; LB14; SOM10; SKC09; YL15], similar to the definitions in the general community detection case (refer to Section 2.2.1). 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 [YL15; JSF+07]. Functional communities have been also suggested to require an intermediate social object that serves to connect people together to truly become social, otherwise they loose interest to remain in contact [Eng05].

For microblogging social networks, a number of works have been published in the field of community detection and its applications. These works can be organised by the subject of community detection that they focus for investigation: (1) early work, (2) structural properties, (3) community properties, (4) real-world events, and (5) community dynamics.

**Table 2.1:** Summary of community detection approaches reviewed for general complex networks.

| Type | Evaluation | Method | Reference | Features |
|---|---|---|---|---|
| Static | Synthetic | l−partition | [CK01] | Minimisation of Edges that cross Communities |
| Static | Synthetic Real-World | Centrality | [GN02] | Centrality Indices to find Community Boundaries |
| Static | Synthetic Real-World | Betweenness | [NG04] | Hierarchical, Iterative Algorithm with Strength Measure |
| Static | Synthetic Real-World | Divisive | [RCC+04] | Quantitative Aspects in Community Definitions |
| Static | Real-World | CPM | [PDF+05] | Statistical Features of Overlapping Communities |
| Static | Synthetic Real-World | Topology | [GA05] | Classification of Communities into Roles |
| Dynamic | Real-World | Topology | [DFC05] | Community Formation during Crisis Periods |
| Static | Real-World | InfoMap | [RB08] | Detection for Weighted and Directed Networks |
| Static | Synthetic | Modularity | [LFR08] | Generation of Artificial Networks for Evaluation |
| Dynamic | Synthetic Real-World | Multiple | [POM09] | Connection of Community Detection to Applications |
| Static | Synthetic Real-World | Multiple | [For10] | Focus on Statistical Methods, Evaluation and Applications |
| Static | Synthetic Real-World | Multiple | [MV13] | Community Detection for Directed Networks |
| Static | Real-World | Multiple | [HBG+14] | Focus on Large-scale Networks and Overlapping Communities |
| Static | Real-World | Multiple | [LC14] | Focus on Large-scale Networks and Evaluation |
| Static | Real-World | Spectral | [YL15] | Focus on Community Definitions and Evaluation |
| Static | Synthetic Real-World | Multiple | [FH16] | Focus on Community Definitions and Evaluation |
| Dynamic | Real-World | Modularity InfoMap | [MDC+16] | Introduces Privacy-preserving Community Detection |
| Static | Synthetic | LFR | [PZ17] | Introduces Naïve Scale-Free Community Detection |

Early work on community detection for microblogging can be found in [JSF+07; BBB+08; Jav08; RHM08; JSF+09]. Java, Song, Finin, et al. were one of the first authors to observe topological properties of microblogging for the purpose of describing communities shortly after the popularisation of the medium [JSF+07]. In this work, a Twitter dataset consisting of 1.4 million posts was collected including follower and friendship networks. It was found that people mostly used the platform to talk about their daily activities and to seek the sharing of information. Furthermore, a taxonomy of intentions is proposed that characterises what motivation users might have for creating content in microblogging. These user intentions were aggregated and compared to implicit communities mined using the Clique Percolation Method (CPM) [PDF+05] over the follower and friendship networks. The authors conclude that communities can be found by considering a set of user intention for composing tweets. Java further experimented with different novel algorithms for identifying communities based on the scale-free properties of the microblogging social network [Jav08]. In this thesis, abstract objects with embedded social functions are considered instead of user intentions for the purpose of defining communities. Moreover, user interaction networks are proposed instead of followers or friendship networks.

Other early works investigated emergent user practices and community awareness in microblogging. Barkhuus, Brown, Bell, et al. studied how users change and share their status within social groups [BBB+08]. In their work, they used a mobile presence application that automatically updated the user location and allowed to tag areas for others to see. This scenario created an awareness effect on the users and it was observed that groups shared ongoing conversations related to the tagged locations. Ryan, Hazlewood, and Makice installed displays called *Twitterspaces* at different locations in community centres with the purpose of studying the encouragement of community awareness and engagement [RHM08]. The screens were designed to retrieved recent tweets from community members and then place them in public spaces, forming the concept of *community at a glance*. Their experiment observed an increase of posting activity within the community and measured an increase in the awareness among community members in the form of birthday reminders, small conversations and organisation to attend community events. In this thesis, these user practices and the location awareness are abstracted into functional social objects that can be used to form communities in microblogging.

More in-depth studies of community formation in microblogging appeared during 2010-2011 and later during 2015-2016 [BGL10; GWT11; DOG15; FH16]. Boyd, Golder, and Lotan examine the mechanism of retweeting (rebroadcasting) in Twitter as a manner for the involved users to form conversations [BGL10]. The act of retweeting can be performed in diverse styles and for a variety of reasons. Following this idea, the authors study how authoring, attribution and

communication quality are established in microblogging. Gruzd, Wellman, and Takhteyev later investigate the feasibility of user communities to actually form in Twitter [GWT11]. They identify that connections in microblogging are asymmetric – following a person does not imply that the person will follow back – and, because of this, these connections depend less on interpersonal interactions. However, the authors argue that Twitter may still have an *implicit sense of community*, i.e. users tend to associate with each other. Using Twitter user networks as a base, the authors show that despite the platform not being designed to support user communities, users still make use of it for forming social connections in an implicit manner. In this thesis, the act of retweeting is leveraged and abstracted into interaction networks for the community detection task.

Identifying communities in online social networks is challenging because there are no universal conventions for the required components [FH16]. For example, there are many interpretations for the definition of community and henceforth there is no universal approach for evaluating community detection algorithms. In addition, the underlying social networks – which can be links between a user and her friends, associates, favourite newscasters, business partners – are often multifaceted. For this reason, Darmon, Omodei, and Garland propose to not ignore this fact and instead incorporate these facets into layered networks for structural community discovery in order to capture valuable viewpoints of the communities [DOG15]. Three types of network connections are considered: activity-based, topic-based and interaction-based. They adopt Twitter as the case study and conclude that the activation of the facets in the network (both how and when) are critical for finding community structures. In this thesis, multi-faceting aspects of the Twitter network are incorporated in the form of user interaction types. In particular, retweets, mentions, quotes and replies are all captured and abstracted for community detection.

Processing microblogging social networks often involves large-scale high data dynamics that should not be disregarded [SLC+12; LB14; AVC+16; ARL18; RPP+17]. Sundaram, Lin, Choudhury, et al. review these dynamic characteristics by linking the social aspects with modern scalable methods [SLC+12]. They observe that real-world communities are based on coherent and sustained interactions (in time), and identify the need for dynamic temporal methods. In this vein, Lu and Brelsford investigate the dynamics of microblogging using the InfoMap community detection method [RB08] in response to extreme events such as natural disasters [LB14]. Using retweets and user mentions networks captured from Twitter, their findings reveal that the community joining and leaving behaviours are not random effects. Users tend to stay in their current state and are less prone to shift communities or join new ones when in solitude. In addition, communities of users quickly shift their topics under the effect of the disaster event.

Additional subjects in relation to community dynamics are the study of user roles [AVC+16] and multiple layers in the social network [ARL18]. Amor, Vuik, Callahan, et al. propose graph-based methods to process Twitter data captured in the context of political debates around a controversial public matter [AVC+16]. In their work, they state that directionality – information in Twitter often flows from the followed to the followers – is a critical aspect in the Twitter social graph and is often ignored in network analysis. The authors propose methods based on random walks and role-based similarity that considers directionality and are able to discover communities that correlate with the observed information flow. In this context, their method is able to identify and classify users by their roles in orchestrating the flow of information across the network. Regarding layered networks (where links are of multiple kind), Aslak, Rosvall, and Lehmann present an approach to enhance InfoMap by estimating layer dependencies at the node-level [ARL18]. When constraining the network using these estimates, InfoMap is able to better uncover user communities with nodes in multiple groups. Therefore, associating network layers at the node-level improves the InfoMap diffusion model for highly dynamic microblogging social networks, which in turn improves the detection of intermittent communities.

Lastly, Rossetti, Pappalardo, Pedreschi, et al. propose an approach named *TILES* for detecting overlapping communities in highly dynamic social networks and tracking their evolution in time [RPP+17]. This algorithm follows an iterative cascading effect strategy, dynamically recomputing community memberships every time a new interaction occurs in the network. The approach is evaluated against other community detection algorithms on networks with annotated ground-truth community structures. The studied networks are both, synthetic (using LFR, Section 2.1.3) and real-world social networks that include user mentions networks from Weibo, the Chinese counterpart of Twitter. The proposed algorithm is demonstrated to be able to achieve better alignment to the ground-truth data than comparable methods and also the computing time is shorter. Moreover, the authors also characterise the properties of the identified communities. In this thesis, the user interaction dynamics in microblogging social networks are investigated. In particular, highly active portions in the lifetime of the community structures in the network are identified to generate temporal sub-communities with better structural quality.

Research on practical applications specifically for community detection in microblogging can also be found [CRF+11; GJK12; BLL15]. Conover, Ratkiewicz, Francisco, et al. propose to investigate how microblogging social media can facilitate communication between communities with different political orientations [CRF+11]. Two social networks of politics-related tweets captured from Twitter regarding the 2010 congressional midterm elections in the United States are considered. The work uses label propagation community detection [RAK07] applied to retweets

interactions to reveal the existence of a polarising bipartite grouping aligned to the two involved political parties. However, this was not the case for user mentions interactions in the same network under study, highlighting again the importance of multi-faceted networks. Instead, in the user mentions network the discovered communities were politically heterogeneous and users with opposing ideologies had considerable higher rates of interaction compared to the retweets case. The explanatory hypothesis is that politically motivated individuals provoke interaction by fuelling divisive content to audiences consisting of ideologically-opposed users.

Another application for community detection in microblogging is natural disaster management. Gupta, Joshi, and Kumaraguru analysed three major disaster events during 2011 – hurricane Irene, riots in England, and earthquake in Virginia – to identify and characterise user communities during the hardship of these events [GJK12]. The motivation in this use case lies in identifying central users that can aid and provide support during the crisis. Their work leverages Twitter social networks to define similarity metrics between the users based on their connectivity and published content. Afterwards, spectral clustering [Lux07] is applied to obtain user communities, then degree centrality is used to identify the central users inside these communities. The detected central users align to the topics and opinions of all the users in the community with high accuracy (80 %), giving way to a representative user role which in turn can be used to summarise the entire community. In another work, Bakillah, R.-Y. Li, and Liang investigate graph clustering techniques for identifying communities in specific physical locations during disaster situations [BLL15]. The algorithm is a fast-greedy optimisation of modularity (FGM) [CNM04] enhanced with semantic similarity for handling the complexities of Twitter social networks. The approach can identify spatial clusters at different points in time. In this thesis, three applications for community detection in microblogging are proposed for two classes of users: end-users and decision makers. The first class aims for regular users of the microblogging service, while the latter aims for specialised users focused on virtual community management.

A summary of the reviewed literature for community detection in microblogging social networks can be found in Table 2.2. The detection methods and network data they use are also observed. As reviewed in this section, the community detection task for microblogging is mostly addressed by means of exploiting static networks, e.g. followers, captured in snapshots [JSF+07; GJK12; LB14; SOM10; SKC09]. However, in this thesis the goal is instead to understand how user communities can form solely through their public user interactions represented as a dynamic network. Furthermore, community detection can be performed with a combination of static networks [DOG15; BLL15; AVC+16], however it is argued that such static networks are expensive to retrieve and maintain fresh in comparison to a real-time stream of messages [DOG15].

**Table 2.2:** Summary of community detection approaches reviewed for microblogging social networks.

| Type | Networks | Method | Reference | Features |
|---|---|---|---|---|
| Static | Followers | CPM | [JSF+07] | Topological Properties and Taxonomy of Intentions |
| Static | Co-location | Analysis | [BBB+08] | Focus on Location Awareness |
| Static | Followers | Multiple | [Jav08] | Based on Scale-free Distribution of Node Degrees |
| Static | Co-location | Analysis | [RHM08] | Focus on Community Awareness and Engagement |
| Static | Retweets | Statistics | [BGL10] | Study of Retweet Behaviour |
| Static | Followers | Analysis | [GWT11] | Topological Properties |
| Static | Multiple | Facets | [DOG15] | Activity, Topic and Interaction Network Layers |
| Static | Retweets/Mentions | Label Prop | [CRF+11] | Application in Politics |
| Static | Followers/Retweets | Spectral Cl | [GJK12] | Similarity Metrics Based on Connectivity |
| Static | Co-location | FGM | [BLL15] | Semantic Similarity |
| Dynamic | Interactions | Multiple | [SLC+12] | Social Aspects of Communities |
| Dynamic | Retweets/Mentions | InfoMap | [LB14] | Community Joining and Leaving Behaviour, Evolutionary Events |
| Dynamic | Followers/Retweets | Flow Prop | [AVC+16] | Random Walks and Role-based Similarity |
| Dynamic | Retweets/Mentions | InfoMap+ | [ARL18] | Layered Networks |
| Dynamic | Mentions | Label Prop | [RPP+17] | Overlapping Communities, Iterative, Evolutionary Events |

For this thesis, a new set of real-world datasets was assembled from Twitter (refer to Chapter 3) instead of attempting to reuse datasets from previous work. This is due to the investigation of all types of interactions available in microblogging, i.e. replying, rebroadcasting, mentioning and quoting, that are not available at the same time in the same datasets in the literature. Furthermore, the Twitter developer terms and data protection policies do not permit to disclose any data directly containing public content[3], including user interactions.

## 2.3 SOCIAL CHARACTERISTICS OF MICROBLOGGING

In this section, the qualities and value of microblogging – represented by Twitter – as a social medium for the formation of virtual communities are explored. Unlike other social media, mi-

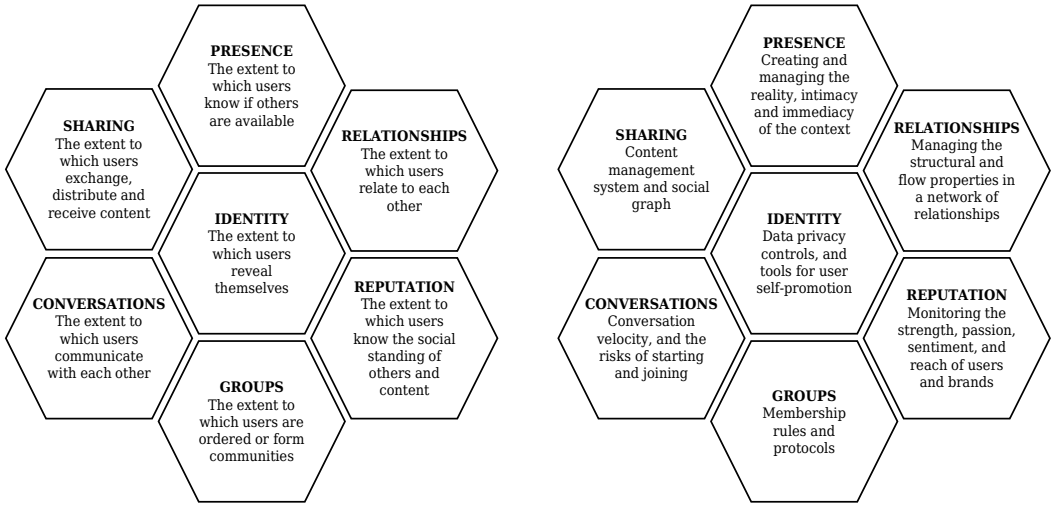
3 Section VII in https://developer.twitter.com/en/developer-terms/agreement-and-policy.html

croblogging social networks have specific characteristics, e.g. high sparsity and fast-pacing, that make them challenging for directly applying structural community detection approaches. Therefore, the social functionalities for different social media such as Facebook, YouTube or LinkedIn will be reviewed and compared to the microblogging case to motivate the need to refine the problem of community detection more precisely for this particular medium. For a complete history of many microblogging services, including their characteristics, refer to Appendix A.

In Twitter, relationships between users are less personal, promoting instead a more open ambient for socialising [SKC09]. 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. Users can choose to follow other users for up-to-date content, often with no approval needed or without requiring to be followed back. However, only 20 % of users tend to follow each other [KLP+10], i.e. a low reciprocity, in contrast to other services such as Flickr (70 % [CMG09]) or Yahoo! 360 (80 % [KNT10]). This suggests that Twitter followers networks might not be adequate for structural community detection. Furthermore, information in Twitter spreads less than five hops away, shorter than in other known social networks [KLP+10], highlighting the strength of microblogging as a medium for rapid information diffusion compared to platforms focused on verified relationships.

Microblogging social networks also differentiate from classic human social networks [NP03]. For instance, the distribution of subscribers is not power-law, the degree of separation is shorter and most links between its users are not reciprocated [KLP+10]. However, Twitter still evidences degrees of homophily (refer to Section 2.1.2). Contact between similar people occurs at a higher rate than among dissimilar members, resembling the behaviour of communities. In contrast to online social networks such as forums, microblogging was not designed explicitly for organised discussions in threads [STA07]. Instead, users can post content more openly without the need to always receive replies or retweets. Nevertheless, forms of discussions are still possible in microblogging that could further lead to the discovery of implicit communities.

To further understand the social capabilities of microblogging social media, a comparison to other services is investigated and the approach in [KHM+11] is adopted. Kietzmann, Hermkens, McCarthy, et al. propose that social networks can be categorised according to the degree of social functionalities they offer: identity, conversations, sharing, presence, relationships, reputation and groups. Therefore, a functional framework can be constructed for characterising social media in terms of these seven core social functionalities – seen in Figure 2.1a – and their implications for businesses, organisations or social community owners – seen in Figure 2.1b.




**(a)** Social Functionalities **(b)** Implications of Functionalities

**Figure 2.1:** The honeycomb of social media functionalities and their implications to business, organisations and community owners, from Kietzmann, Hermkens, McCarthy, et al. [KHM+11]

Different social media services can implement a number of these social functionalities at different levels. Depending on the level of the implemented features, social media services can support varied strategies for community managers to meet their particular needs. Because of this diversity, large companies tend to have presence in multiple social networks at the same time. For instance, the practical applications in this thesis (refer to Chapter 6) are designed to provide user communities not only to end-users but also to decision makers.

An important observation in [KHM+11] is that online social networks tend to concentrate on three to four of the blocks of the honeycomb framework in Figure 2.1. Borrowed from that work, this is illustrated in Figure 2.2 for four well-known social media services (green in the Figure): Facebook, Foursquare, LinkedIn and YouTube. Their different social functionalities and their level of implementation – ranging from 0 (none) to 2 (dominant) – can be seen. Facebook relies on strong reciprocal friendship relationships, FourSquare focuses on physical user presence, LinkedIn greatly values user identity and YouTube concentrates on video content sharing. In this thesis, the social functionalities of the Twitter microblogging service (blue in the Figure) were researched for comparison, and the results are detailed in Section 2.3.1.

Kietzmann, Hermkens, McCarthy, et al. further studied different audiences and the social media characteristics that they prefer. In Table 2.3, the first four columns summarise their original findings in terms of social functionalities available for the intended audience. In this thesis, the interest is on community functionalities and microblogging social networks, which were not considered in the original work. Therefore, community functionalities were investigated and added as complement in the last row and fifth column of Table 2.3 for comparison.

| | Facebook | Foursquare | LinkedIn | YouTube | Twitter |
|---|---|---|---|---|---|
| Presence | 1 | 2 | 0 | 0 | 0 |
| Sharing | 0 | 0 | 0 | 2 | 1 |
| Relationships | 2 | 1 | 1 | 0 | 1 |
| Identity | 1 | 1 | 2 | 0 | 1 |
| Conversations | 1 | 0 | 0 | 1 | 2 |
| Reputation | 1 | 0 | 1 | 1 | 1 |
| Groups | 0 | 0 | 0 | 1 | 0 |

Figure 2.2: Social functionalities of popular online social networks and their level of implementation, ranging from 0 (none) to 2 (dominant), as reported by Kietzmann, Hermkens, McCarthy, et al. [KHM+11]. The data for the Twitter microblogging service is an extension from this thesis.

Table 2.3: Social media audiences and the characteristics that online social networks offer to them, partially from Kietzmann, Hermkens, McCarthy, et al. [KHM+11]. The data for microbloggers audience and the community column are an extension from this thesis.

| Audience | Examples | Content Type | Reputation | Community |
|---|---|---|---|---|
| General Masses | Friendster, Hi5, Facebook, Reddit, Boards | Posts, Notes, Stories | Likes, Subscribers | Explicit Forums and Groups, Explicit Replies |
| Professionals | LinkedIn | Careers | Connections | Explicit Groups |
| Media Sharing | MySpace, YouTube, Flickr, SoundCloud | Video, Images, Music | View Counts, Likes, Subscribers | Comments and Shared Playlists |
| Bloggers | Gizmodo, Tech Crunch, TMZ, Individuals | Posts, Articles | Subscribers, Likes | Comments |
| Journalists and Collectors | Digg, Delicious | Links, Websites | Likes, Ratings | Comments and Explicit Replies |
| Microbloggers | Twitter, Foursquare | Statuses, Check-ins | Followers, Likes, Ratings | Explicit Replies, Rebroadcasting, Users Lists |

### 2.3.1 Social Functionalities of Microblogging

Microblogging social networks were not considered in the work from Kietzmann, Hermkens, McCarthy, et al. [KHM+11] and therefore not contrasted to the other social media services in the study. However, with the honeycomb framework it is possible to quantify what social functionalities are provided by microblogging social media and at which levels. Afterwards, this characterisation can be compared against the other types of social media and explore how they differ. Therefore, the following assessment of the seven social functionalities of the honeycomb proposed in [KHM+11] (seen in Figure 2.1) and their level of implementation in microblogging services – represented by Twitter – is proposed below, and also summarised in Figure 2.2.

**PRESENCE** is the extent to which users can know if other users are accessible, i.e. knowing where others are (virtual/real world), and whether they are available. In Twitter, location is vaguely reported by users and there is **no support** for availability (no implementation), in contrast to services such as Foursquare where user presence is openly broadcasted.

**SHARING** is the extent to which users exchange, distribute and receive content. In many cases, sociality is about the objects that mediate the social ties between people and the reason they associate with each other. In some services such as Facebook, sharing functions are more restricted in audience, while in others such as YouTube sharing is fundamental to their social model. In Twitter, shareable objects are Tweets, web links, videos and photos, all of which can be shared via simple means. Implementation level is **medium**.

**RELATIONSHIPS** is the extent to which users can be related to other users, i.e. two or more users have some form of association that leads them to converse, share objects, meet up or simply list each other as a friend or fan. In Twitter there is basic support for followers and users lists, and a simple user-mentioning syntax. However, Twitter is more focused on general ambient awareness than rich inter-personal communication. Implementation level is **low**.

**IDENTITY** is the extent to which users reveal their identities in a social media setting, i.e. disclosing information such as name, age, gender or location. In Twitter there is good support for identity as many users give semi-accurate data, while in other services such as Facebook there is a much stronger validation of user identity. Implementation level is **medium**.

**CONVERSATIONS** is the extent to which users communicate with other users in a social media setting, i.e. how well the platform facilitates communication among individuals and groups. In Twitter there is rich support for conversations, including rebroadcasting (retweets), replies, mentions and content tagging via hashtags. Moreover, because Twitter does not require user registrations to read public conversations, it is considered more open than other services such as Facebook. Implementation level is **high**.

**REPUTATION** is the extent to which users can identify the standing of others, including themselves, in a social media setting. In Twitter, the number of followers is often used for reputation, however it has limited value because it correlates more to the popularity of users, rather than to how many of them actually consume their content. Moreover, the number of followers does not relate to the quality of the content generated by the user. In contrast, other services such as Facebook or YouTube offer better reputation functions such as Like/Dislike buttons. Implementation level is **low**.

**GROUPS** is the extent to which users can form communities and sub-communities. The more social a network becomes, the bigger will be the group of friends, followers and contacts. In Twitter there is **no support** for rich user grouping besides user lists. Moreover, followers are not required to be followed back, preventing the formation of reciprocal groups. Followers networks in microblogging are not designed for community functionalities.

The above social functionalities and and their identified implementation levels for Twitter allow to construct a representation of the honeycomb framework as shown in Figure 2.2. In summary, Twitter is strongly designed for ambient awareness and therefore is more focused on open conversations. Twitter also mildly values identity, hence it is unable to value strong relationships [KHM+11]. Regarding reputation, Twitter offers basic tools for social positioning, i.e. reputation is based on the number of followers of a user. This approach for reputation is weak and a stronger method would be the accounting of user mentioning. The best approximation to groups in Twitter are user lists, which are limited. In contrast, Facebook groups are explicitly managed with a rich set of tools. Other social media types such as blogs facilitate rich and long conversations, Twitter instead focuses on fast-pacing and real-time delivery of short content.

Being a truly open medium for users to express themselves freely without the need of replies, Twitter is not well suited for strongly managed communities, i.e. there is weak bi-directional conversation. Moreover, Twitter users tend to share information more than engaging in long discussions. Nevertheless, implicit community structures around social objects such as real-world events, films, music or news might be present in Twitter user interactions and could potentially be discovered. In this thesis, finding implicit communities is regarded as an important social functionality for assisting users to make sense of microblogging content and therefore minimise their effort for finding interesting publications.

## 2.4 POSITION OF THE THESIS

The research methodology used in this thesis is partly inspired on the previous work of J. Yang and Leskovec [YL15]. The authors empirically study a set of structural community definitions over more classical online social networks. However, this thesis differentiates from their work in that: (1) the original study is extended from traditional online social networks to the microblogging case, taking into account the particularities of the platform such as its temporal dynamics, (2) the original experiments are adapted to address the challenges imposed by microblogging

such as its data volume and sparsity, and (3) new experiments are proposed to characterise and model the time dynamics of the microblogging scenario for community detection.

In general, the proposed research questions and sub-questions in the thesis contemplate the analysis of microblogging social networks at the meso and macro level, where these networks are investigated considering them as randomly distributed, scale-free and complex networks. For this, thirteen different structural community measures [YL15], i.e. community definitions, are investigated explicitly for the case of microblogging social networks.

To complement, the social functionalities of microblogging services (represented by Twitter) are also investigated in this chapter for comparison to other social services such as Facebook, LinkedIn or YouTube (refer to Section 2.3.1). The interest of this thesis is on the community functionalities and microblogging social networks which were not considered in the related work from Kietzmann, Hermkens, McCarthy, et al. [KHM+11]. Finding implicit communities is regarded as an important social functionality in this thesis for assisting users to make sense of microblogging content and therefore minimise their effort for finding interesting content.

The position of the thesis in relation to different areas of the problem of community detection for the case of microblogging social networks can be found in the next sections.

### 2.4.1 Assembling Microblogging Ground-Truth Data

In contrast to the related work discussed in this chapter, in this thesis ground-truth user communities are defined through explicitly labelled social functions in microblogging social media such as topic tags, locations or external users referenced in common (refer to Chapter 3). Furthermore, the problem of community detection is defined through structural scoring functions that measure different structural aspects of communities such as conductance, clique formation or modularity. This provides the necessary flexibility to study microblogging social streams.

In the literature, the problem of community detection evaluation in complex networks have been extensively studied and mostly focus on using synthetic and manually annotated real-world data [LFR08; MV13; HBG+14; LC14; YL15; FH16]. In this thesis, evaluating community detection is defined as quantifying the ability of the structural definitions to model functional ground-truth communities. These ground-truth communities are assembled from explicitly-labelled social functions in real-world large-scale streams of microblogging data, independently of any community detection method, ensuring their validity for evaluation.

### 2.4.2 Defining Microblogging Communities

This thesis gives particular attention to the social functions that define user communities in microblogging social networks by investigating their alignment to their underlying user interaction topologies. Early work on community detection for microblogging evidences the usage of abstract objects to model sociality [JSF+07; BBB+08; Jav08; RHM08; JSF+09]. In this thesis, abstract objects with embedded social functions are also considered and preferred instead of user intentions for the purpose of defining communities in microblogging social media (refer to Chapter 3). Furthermore, social user practices and location awareness are abstracted into functional social objects that can be used to form communities in microblogging.

### 2.4.3 Defining Microblogging Social Networks

In this thesis, user interaction networks in microblogging are proposed instead of followers or friendship networks used in other works as reviewed in this chapter. In-depth studies of community formation in microblogging examine the mechanism of retweeting (rebroadcasting) in Twitter as a manner for the involved users to form conversations [BGL10; GWT11; DOG15; FH16]. In this thesis, this is further expanded using the acts of retweeting, quoting, replying and mentioning other users. Furthermore, user interaction dynamics in microblogging social networks are also investigated. In particular, highly active portions in the lifetime of the community structures in the network are identified to generate temporal sub-communities with better structural quality. In the literature, the community detection task for microblogging is mostly addressed by means of exploiting static networks, e.g. followers, captured in snapshots [JSF+07; GJK12; LB14; SOM10; SKC09]. However, in this thesis the goal is instead to understand how user communities can form solely through their public user interactions represented as a dynamic network.

### 2.4.4 Practical Applications in Microblogging

Three applications for community detection in microblogging are proposed in this thesis for two classes of users: end-users and decision makers. The first class aims for regular users of the microblogging service, while the latter aims for specialised users focused on virtual community management. The practical applications in this thesis (refer to Chapter 6) are designed to provide user communities not only to end-users but also to decision makers.

## 2.5 CHAPTER SUMMARY

In this chapter, the foundations and position of the thesis were discussed. The fundamental concepts used in the thesis were introduced and discussed: social networks (Section 2.1.1, user communities (Section 2.1.2 and community detection (Section 2.1.3). Furthermore, a systematic review of the recent literature involving the problem of community detection in general (Section 2.2.1) and in the context of microblogging social networks (Section 2.2.2) was presented. The social functionalities of microblogging services that motivate the interest for investigating community detection in this particular medium were discussed (Section 2.3). Lastly, the position of the thesis in the literature discussed and summarised (Section 2.4).

The background and literature review in this chapter aimed to support the research questions, sub-questions and outcomes of this thesis – introduced in Chapter 1 – which focus on the investigation of the problem of community detection, the characterisation of user communities, the evaluation of different detection approaches and the improvement of existing techniques in the context of microblogging social networks.

# 3 BUILDING COMMUNITIES IN MICROBLOGGING

In this chapter, the construction of ground-truth functional communities and user interactions networks from microblogging social media – represented by Twitter – is studied. The following main research question, proposed for this preliminary stage in Chapter 1, is addressed.

(RQ1) → *How can microblogging ground-truth and structural data for community detection be assembled and modelled in Twitter social streams?*

To provide an answer to this research question, the following research sub-questions are also proposed for this stage, and are investigated in detail in this chapter.

- (RQ1.1) → *For constructing ground-truth data, how can independent, explicitly user-labelled, functional communities be modelled in Twitter social streams?*
- (RQ1.2) → *For constructing structural data associated to the ground-truth in (RQ1.1), how can networks of user interactions be modelled in Twitter social streams?*
- (RQ1.3) → *What are the global properties, e.g. size and membership distributions, of the defined ground-truth functional communities in (RQ1.1) and (RQ1.2)?*

These research sub-questions provide the following contributions: (1) a working definition of functional communities for microblogging, (2) a methodology for building ground-truth functional communities from microblogging user interactions designed for, but not limited to, streaming Twitter data, and (3) a characterisation and understanding of the global properties of functional communities in microblogging social media.

First the working definitions of functional and structural communities used in this thesis will be formally introduced and discussed. Afterwards, the approach for building functional communities to be used as ground-truth is presented, followed by the method for building the graph-based model of live user interactions, both for microblogging. The real-world experimental datasets built using the above methodology to be used in this thesis will be introduced, and their global properties will be investigated.

## 3.1 DEFINING FUNCTIONAL AND STRUCTURAL COMMUNITIES

User communities are traditionally seen as groups of vertices in a network that have higher chances of being connected to each other than to other vertices, however other definitions are also possible [For10]. Community detection is a challenging subject because there are many valid interpretation of its aspects, e.g. the definition of community itself, therefore validation and comparison of approaches in various domains is difficult. Therefore, when conducting community detection analysis, it is necessary to consider what will be the *characteristic* to be used to group users together in communities for the particular case under study. Moreover, in order to evaluate the performance of a set of detection algorithms, the chosen community definition must be comparable among all of the approaches in the evaluation.

The definition of a user community in Twitter – and therefore for microblogging – is often described in the literature as a group of nodes more densely connected to each other than to nodes outside the group [TL10; JSF+07; GJK12; LB14; SOM10; SKC09; YL15]. However, it is argued in this thesis that, in microblogging social networks, people do not seek to be closely related but instead are more curious about the collective public opinion of the global user-base, unlike other social platforms where strong relationships between users are preferred, e.g. Facebook.

Instead of attempting to craft yet another community definition specifically for microblogging, this thesis will prefer and evaluate a more flexible open and wider interpretation of user communities. In particular, a differentiation between *functional* and *structural* definitions of user communities will be adopted [YL15], as seen in Figure 3.1 and described below.

**FUNCTIONAL COMMUNITIES** are defined as groups of users sharing a common and independent social *function*, e.g. fans of the same football team, people living in the same area or discussing the same topic. Social functions such as topic tags, locations or external users referenced in common are independent from any underlying social network, and instead are explicitly stated by the users in their messages.

**STRUCTURAL COMMUNITIES** are defined as groups of users with a particular pattern in their *connectivity in a network*, e.g. their average node degree or clustering coefficient.

It is then argued that functional communities can be uncovered from structural patterns in a network of live interactions. Moreover, functional communities will represent ground-truth information because users themselves explicitly state the social function they use in their posts, e.g. referencing the same hashtag or mentioning the same celebrity.

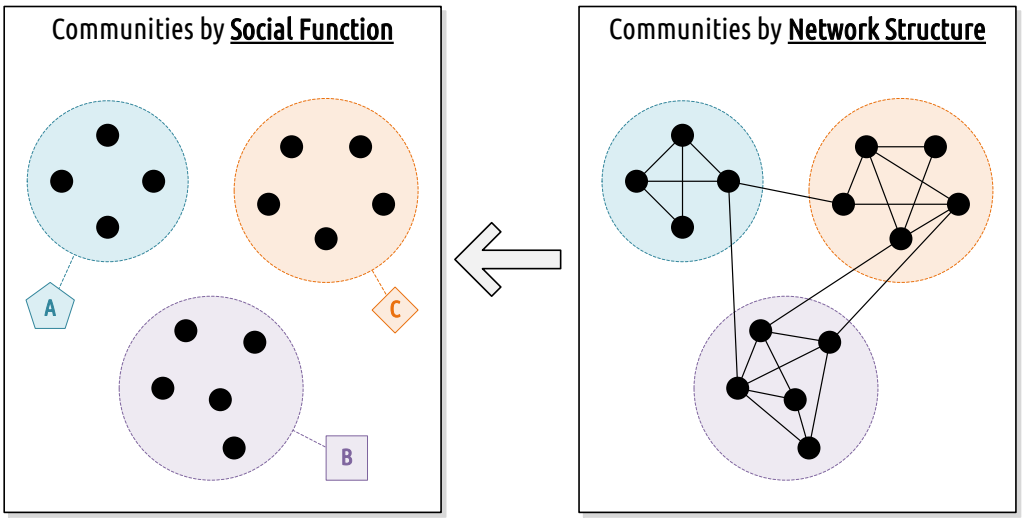



**Figure 3.1:** Functional and structural community definitions. In *functional* communities (left), users are grouped according to specific social functions (A, B, C) such as belonging to the same country, liking the same artist or attending the same class. In *structural* communities (right), users are grouped according to defined network patterns, e.g. groups with high average node degree.

This thesis investigates the relationship between these two types of definitions of communities, considering the task of community detection as the discovery of user groups based on a structural definition that later correspond to ground-truth functional groups [YL15]. Moreover, the task of community detection in microblogging is evaluated in terms of the quality of the alignment between connectivity patterns in the interactions network, as described in Chapter 4, and the explicitly labelled social functions given by the users in the ground-truth.

## 3.2 GROUND-TRUTH FUNCTIONAL COMMUNITIES

In microblogging, users create short messages that are rapidly spread among their followers. These messages can contain a number of social functions (activities around different social objects [Eng05]) that the users themselves assign to them. In Twitter, these messages are called *Tweets* and an example post using different social functions can be seen in Figure 3.2. In this example, a user *John M. Doe (@johndoe)* posted a message replying to another user with the handle *@janedoe*. Every message in Twitter is time-stamped, e.g. *Aug 28, 2018 at 2:10pm*. In his reply, John Doe highlighted some of his feelings about the weather using *hashtags*, e.g. *#lovely* and *#sunny*. Hashtags in Twitter are easily searchable through the user interface and if enough users employ the same, they can become global trending topics. Furthermore, John Doe also provides Jane Doe with a web link to the weather forecast, and mentions a third user (*@goober*) so he also

is aware of the message. When a user is mentioned in Twitter, the user interface notifies her of the mention and the Tweets where she appears. John Doe also chose to disclose his current location at the time of posting the Tweet, e.g. *Galway, Ireland*. Finally, Twitter provides options to rebroadcast or quote Tweets using special *Retweet* and *Quote* buttons in the user interface. Retweets are one of the fundamental mechanisms for information spread in Twitter [BGL10]. Quotes are similar to Retweets but are allowed to have additional content.

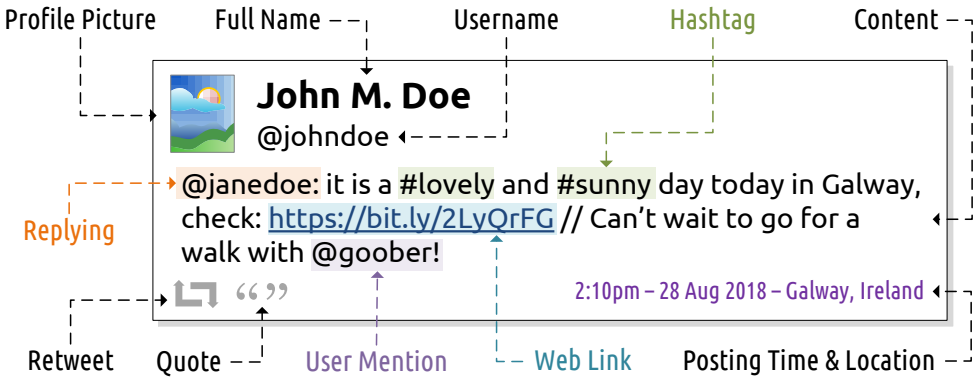



**Figure 3.2:** An example Tweet from the Twitter microblogging service. A number of *social functions* that users themselves can assign to their messages can be identified: replying, mentioning, retweeting, quoting, web resource linking, tagging and location.

All of the above activities are always associated with a *social object* in Twitter. A social object is the element that connects people together to truly become social, otherwise they lose interest [Eng05]. For example, when the reply, retweet, quote or mention activities are used, the associated social object is the user being replied to, retweeted, quoted or mentioned. Similarly, when using the tagging or linking activities, the social object becomes the hashtag and the web link, representing a topic or resource that the users are interested on. All these social activities are supported by Twitter as features in its software platform (refer to Section 2.3.1). Therefore, **ground-truth functional communities** in Twitter can be built 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 instance, if a set of users $\{u_1, u_2, u_3\}$ use the same hashtag $h$ (their social object), then a ground-truth community $C_h = \{u_1, u_2, u_3\}$ after $h$ is created.

The following eight social functions and social objects available in Twitter are considered in this thesis for building different microblogging ground-truth functional community types.

**HASHTAGS** are the social objects used by the tagging social function and represent topics of interest for Twitter users. A functional community of type *hashtags* is formed when a group of users use the same hashtag social object, e.g. a trending topic or event of interest.

**MENTIONS** is the social function where a user $u_i$ mentions another user $u_j$. A functional community of type *mentions* is formed when a group of users mention the same *external user*, e.g. a celebrity, thus the celebrity constitutes the social object for the community members.

**RETWEETS** is the social function where a user $u_i$ rebroadcasts (retweets) the content of another user $u_j$. A functional community of type *retweets* is formed when a group of users retweet the same *external user*, e.g. a newscaster posting shocking news. The retweeted external user then becomes the social object for the community members.

**QUOTES** is the social function where a user $u_i$ quotes the content of another user $u_j$. A functional community of type *quotes* is formed when a group of users quote the same *external user*, e.g. provide their opinions over a statement of a politician. The quoted external user then becomes the social object for the community members.

**COUNTRIES, CITIES, PLACES** are social objects that represent physical locations at different granularities. In Twitter, a *place* is an object representing a well-known location that can be embedded in Tweets. Locations can be of any kind, for instance as broad as *"The Promenade"* or as narrow as *"The Cinema"*. Places can also contain country and city attributes, therefore functional communities based on this information can be constructed as well, e.g. the promenade is in Galway or the cinema is in New York. A functional community of type *countries*, *cities* or *places* is formed when a group of users post Tweets from the same location social object, i.e. country, city or place respectively.

**URLS** are social objects that represent a resource, e.g. website, image or video, on the Internet. A functional community of type *urls* is formed when a group of users embed the same URL social object in their Tweets, e.g. a link to an interesting cooking recipe.

Even though some of these social functions are also used to build the interactions network in the next section and could be considered as an inherent bias, it is stressed here that when building the ground-truth functional communities, these functions are always used in context with an external factor, i.e. social object, and not between the community members. For example, when a ground-truth functional community $C_m = \{u_1, u_2, u_3\}$ of type *mentions* with users mentioning the same user $u_m$ is constructed, $u_m$ is not in this community nor interacts with its members. Instead, $u_m$ is considered as an external motive for members of $C_m$ to be connected socially.

Furthermore, two simple build restrictions are imposed during construction for each ground-truth functional community: (1) each group must have at least three members to facilitate the study of community scoring functions based on clique structures, and (2) communities with

more than one connected component in the underlying user interactions network must be separated and each component treated as an independent ground-truth community.

Users participating in functional communities may not be completely aware of the social objects creating connections between through their social functions. Therefore, it is further investigated in this thesis if such connection really exists through their live user interactions.

## 3.3 USER INTERACTIONS NETWORK

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, retweeting and quoting, can be used to form a network of interactions between users [HH14].

Based on [HPH+15; YLL+14], four concrete types of Twitter interactions are considered for building a network of live user interactions from a stream of Tweets in this thesis: mentions, quotes, replies and retweets. These do not refer to the social functions with the same names in Section 3.2, but to interactions that each user perform towards other users directly. In Twitter, replies, retweets and quotes are also reported as user mentions due to their syntax in the content. However, in this work they are all distinguished, i.e. a replied, retweeted or quoted user is not considered as a mentioned user. Furthermore, and for simplicity, the interactions network will be considered as undirected. Therefore, an undirected weighted network $G = (V, E, W)$ is proposed with a set of user vertices $u \in V$, interaction edges $e = (u_i, u_j) \in E$ and time-aware, typed, edge weights $w(e, t, type) \in W$. The possible types for each edge are $type \in \{mentions, quotes, replies, retweets\}$. At every time $t$ a user $u_i$ interacts with another user $u_j$ using any of the defined interaction types in $T$, an edge $e = (u_i, u_j)$ is created in the network and the edge weight $w(e, t, type)$ is incremented by one.

To alleviate excessive granularity in the temporal weights $w(e, t, type) \in W$ during network construction, the observed times $t$ are grouped by applying a quantisation function $Q(t, q) = \lfloor t/q \rceil \cdot q$, where $\lfloor \cdot \rceil$ is the nearest integer operator and $q$ is a quantisation value in the same unit as $t$. For example, if $t$ is measured in seconds, edge weights can be binned per minute ($q = 60$), hour ($q = 3600$), day ($q = 86400$) or any arbitrary quantisation $q$ required by the researcher. The proposed quantisation reduces the time resolution of the community dynamics, however it also reduces the noise in the data by aggregating edges into more meaningful time scales.

The proposed network model with an example can be seen in Figure 3.3 and the complete procedure for building this network from a stream of Tweets is presented in Algorithm 3.1.

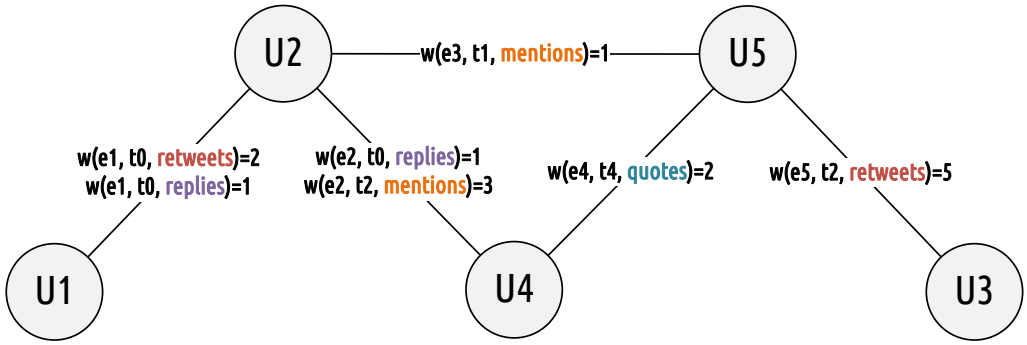



**Figure 3.3:** Proposed undirected user interactions network model for Twitter based on its available interaction types, e.g. retweets, replies or mentions. For each interaction edge $e = (u_i, u_j) \in E$ of $type \in T$, a weight $w(e, t, type) \in W$ is recorded for every occurrence time $t$, with a quantisation $t = Q(t, q)$ applied to it. In this example, five edges $\{e_1, \ldots, e_5\} \in E$ are shown.

In the example in Figure 3.3, five users $\{u_1, \ldots, u_5\} \in V$ are modelled in the microblogging social network $G = (V, E, W)$. Among them, the five interactions $\{e_1, \ldots, e_5\} \in E$ described below have occurred (note that times can be sparse, i.e. there is no $t_3$).

- $t_0$: 2 *retweets* and 1 *reply* between $u_1$ and $u_2$ ($e_1$), 1 *reply* between $u_2$ and $u_4$ ($e_2$).
- $t_1$: 1 *mention* between $u_2$ and $u_5$ ($e_3$).
- $t_2$: 3 *mentions* between $u_2$ and $u_4$ ($e_2$), 5 *retweets* between $u_3$ and $u_5$ ($e_5$).
- $t_4$: 2 *quotes* between $u_4$ and $u_5$ ($e_4$).

Initially, building separate networks for each interaction $type \in T$ was under consideration. However, a pair-wise network overlap analysis in the experimental data – as described later in Section 3.4 – revealed a low average edges overlap of ≈2.39 %. With this result it is concluded that combining all the interaction types in a single network generates a richer structure overall while maintaining the simplicity of the model.

## 3.4 EXPERIMENTAL DATASETS

In this thesis, real-world Twitter data streams under different settings and periods of time are investigated. The Twitter Streaming API offers two modes for collection of data: the *filter* and the *sample* endpoints [MPL+13]. The first can retrieve streams using defined keywords, geographical coordinates and users to follow, while the latter provides a global unspecified random sampling (estimated to ≈1.5 % by Morstatter, Pfeffer, Liu, et al.) of Tweets being currently published.

**Algorithm 3.1:** Procedure for constructing a weighted social user interactions network from a stream of Tweets, with quantised time-aware and typed edge weights.

**Data:** a stream of Tweets $\mathbb{T} = \{tw_0, tw_1, \ldots, tw_n\}$
**Data:** a time quantisation parameter q in seconds, e.g. 3600
**Result:** an interactions network $G = (V, E, W)$ for $\mathbb{T}$

$V \leftarrow \emptyset$, $E \leftarrow \emptyset$, $W \leftarrow \emptyset$

**Function** `addToNetwork`$(u_i, u_j, time, type)$ **is**
  $e \leftarrow (u_i, u_j)$
  $V \leftarrow V \cup \{u_i, u_j\}$, $E \leftarrow E \cup \{e\}$
  **if** $w(e, t, type) = \emptyset$ **then**
    $w(e, t, type) \leftarrow 0$ `/* initialise weight if edge not yet seen */`
  **end**
  $w(e, t, type) \leftarrow w(e, t, type) + 1$
**end**

`/* The mentions, replyTo, retweetOf and quoteOf functions extract all the mentions, any replied user, any retweeted user and any quoted user from a given Tweet respectively. */`
**for** $tw_i \in \mathbb{T}$ **do**
  $author \leftarrow$ `author`$(tw_i)$
  $time \leftarrow$ `Q(time(`$tw_i$, q`))` `/* quantise the time of posting of the Tweet */`
  **for** $m_i \in$ `mentions`$(tw_i) -$ `replyTo`$(tw_i) -$ `retweetOf`$(tw_i) -$ `quoteOf`$(tw_i)$ **do**
    `addToNetwork`$(author, m_i, time, \{mentions\})$
  **end**
  **if** `replyTo`$(tw_i) \neq \emptyset$ **then**
    `addToNetwork`$(author,$ `replyTo`$(tw_i)$ $, time, \{replies\})$
  **end**
  **if** `retweetOf`$(tw_i) \neq \emptyset$ **then**
    `addToNetwork`$(author,$ `retweetOf`$(tw_i)$ $, time, \{retweets\})$
  **end**
  **if** `quoteOf`$(tw_i) \neq \emptyset$ **then**
    `addToNetwork`$(author,$ `quoteOf`$(tw_i)$ $, time, \{quotes\})$
  **end**
**end**

In total, five streams from Twitter were collected under different settings and periods of time. Using the filter endpoint, 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 were captured. To complement the study, one stream from the sample endpoint was also captured. All the collected datasets are described below and were selected according to different contexts of interest for this thesis: (1) wide audiences in microblogging (many users around few specific topics), (2) temporal periodicity in microblogging streams (users or topics that can re-appear multiple times), (3) topic independent microblogging data (many users around many topics), and (4) random sampled microblogging streams (unfiltered content). All these contexts influence the number, size and structure of the potential ground-truth functional communities that can be constructed from them, and therefore are important to study in this thesis.

#### POPE CONCLAVE 2013

This dataset was captured during the *Catholic Pope Conclave*[1] event in 2013, for the context of a **wide audience** in microblogging. The data spans for ≈2 days and contains 460 334 Tweets and 285 569 users. The filter endpoint was used and configured to listen for event-related hashtags and users to follow such as the hashtags #Conclave, #HabemusPapam, and the Twitter accounts of newscasters and candidate cardinals[2]. In all the chapters of the thesis, this dataset is referred to as Pope2013.

#### POPE CONCLAVE 2013 (SAMPLED)

This dataset was captured in parallel to Pope2013 using the sample endpoint, which returns a random sample of every Tweet in the platform and not only specific posts related to the Pope Conclave event. The data spans for the same days but instead contains 9 904 068 Tweets and 8 787 088 users. This dataset is meant to investigate **random sampled** microblogging data. In all the chapters of the thesis, this dataset is referred to as Pope2013-Spl.

#### FIFA WORLD CUP 2014

This dataset was captured during the *FIFA World Cup*[3] event in 2014, also for the context of a **wide audience** in microblogging, but much larger than Pope2013. The data spans for ≈34 days and contains 27 173 102 Tweets and 8 015 322 users. The filter endpoint was used and configured to listen for event-related hashtags and users to follow such as the hashtags #WorldCup, #Brazil2014, and the Twitter accounts of newscasters and participating football teams[4]. In all the chapters of the thesis, this dataset is referred to as WorldCup2014.

#### RTÉ 2015

RTÉ is the public TV and Radio broadcaster of Ireland. This dataset captures Tweets related to different *TV programmes being broadcasted live by RTÉ* during 2015 and is meant to investigate **temporal periodicity** in microblogging streams. The data spans for ≈63 days and contains 2 065 755 Tweets and 720 954 users. The filter endpoint was used and configured to listen for manually curated hashtags, keywords and users to follow such as #GreysAnatomy and #ThewalkingDead, related to each TV programme. Moreover, the filtering terms were dynamically configured according to the broadcasting time of each TV programme. In all the chapters of the thesis, this dataset is referred to as RTE2015.

1 https://en.wikipedia.org/wiki/2013_papal_conclave
2 https://web.archive.org/web/20130908044018/http://expandedramblings.com/index.php/how-to-follow-the-conclave-to-select-the-next-pope-on-a-mobile-device/
3 https://en.wikipedia.org/wiki/2014_FIFA_World_Cup
4 http://blog.twitter.com/2014/follow-the-worldcup-action-on-twitter

**Table 3.1:** Summary of the constructed Twitter ground-truth datasets for the thesis.

| Dataset | Timespan | Nodes | Edges | Communities |
|---|---|---|---|---|
| Pope2013 | ≈2 days | 238 368 | 303 742 | 11 580 |
| Pope2013-Spl | ≈2 days | 6 593 649 | 6 140 684 | 40 812 |
| WorldCup2014 | ≈34 days | 6 932 106 | 15 854 811 | 361 559 |
| RTE2015 | ≈63 days | 643 292 | 1 446 852 | 56 025 |
| Ireland2017 | ≈245 days | 1 067 982 | 2 826 754 | 62 562 |

**Table 3.2:** Summary of the user activity of the Twitter ground-truth datasets for the thesis. $A_u$ 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, sparse, interactions.

| Dataset | $A_u$ | $A_i$ | $A_c$ |
|---|---|---|---|
| Pope2013 | 2.1082 | 0.6604 | 18.7042 |
| Pope2013-Spl | 4.0370 | 0.5098 | 27.1706 |
| WorldCup2014 | 114.1077 | 32.3110 | 334.9153 |
| RTE2015 | 163.6430 | 84.8706 | 687.7683 |
| Ireland2017 | 1483.4306 | 355.7120 | 3881.9658 |

#### IRELAND 2017

This dataset exclusively captures Tweets with embedded location information using the filter endpoint configured to track location-enabled Tweets in *Ireland* during 2017 using the geo-bounding box $(-10.6696, 51.4199, -5.9947, 55.4351)$. Because no specific hashtags nor users to follow were used, it is meant to investigate microblogging streams in a **topic independent** context. The data spans for ≈245 days and contains 7 699 178 Tweet and 1 086 862 users. In all the chapters of the thesis, this dataset is referred to as Ireland2017.

For each of the above datasets, and using the methodology presented in this chapter, sets of ground-truth functional communities and a network of user interactions were constructed. A summary of the general network properties and ground-truth functional communities is in Table 3.1. To complement this, a summary of the user activity in this ground-truth is in Table 3.2. A total of 532 538 ground-truth functional communities were built from a total of 47 302 437 Tweets. In all the datasets, a quantisation $q = 3600$ was used for storing *hourly* time observations, i.e. all times reported in the thesis have a granularity of one hour. This quantisation was chosen empirically to be sufficiently granular, i.e. 24 data points per day. Note that the average user activity and interaction times are very short for the Pope2013 and Pope2013-Spl datasets.

More detail on the the number of communities constructed for each type and in each dataset can be seen in Table 3.3. The average number of communities per day considering each the timespan of each dataset is also reported. It must be first noted that the *quotes* social function was introduced by Twitter in April 2015 as a *retweet with comment* feature[5]. This is reflected on

5 https://techcrunch.com/2015/04/06/retweetception/

**Table 3.3:** Number of communities constructed per type for the Twitter ground-truth datasets of the thesis.

| | Pope2013 | Pope2013-Spl | WorldCup2014 | RTE2015 | Ireland2017 |
|---|---|---|---|---|---|
| *cities* | 3 | 43 | 193 | 16 | 639 |
| *countries* | 9 | 264 | 396 | 34 | 684 |
| *hashtags* | 6565 | 9313 | 144 507 | 19 883 | 13 457 |
| *mentions* | 1886 | 29 780 | 64 418 | 9418 | 37 359 |
| *places* | 3 | 55 | 278 | 17 | 1608 |
| *quotes* | 0 | 0 | 0 | 641 | 8815 |
| *retweets* | 695 | 906 | 27 295 | 6083 | 0 |
| *urls* | 2419 | 451 | 124 472 | 19 933 | 0 |
| **Total** | 11 580 | 40 812 | 361 559 | 56 025 | 62 562 |
| **Avg/Day** | 5281.0844 | 18 612.4021 | 10 709.7815 | 893.4575 | 255.8448 |

datasets prior to 2015 not having any communities built for the *quotes* social object type. Also important to note is the limitation in the Twitter Streaming API of retweets not being delivered when using location-based filtering. This can be observed in the Ireland2017 dataset, where there were no *retweets* communities constructed. It can be observed that in the same dataset, the *urls* social object also resulted in no communities. However, this is an effect of the minimum of three members condition imposed during functional communities construction. In this case, 2 103 450 functional communities were initially assembled. Then, after separating by connected components, a total of 1 926 451 remained, however all of these communities had less than three members, resulting in zero *urls* communities being retained in the dataset. This observation, together with the relatively low number of communities overall in Ireland2017 (an average of 255.8448 communities per day), highlights the fragmentation and sparsity of microblogging data in the context of a location-based capture method.

Lastly, in Section 3.3, it is proposed that a single interactions network for all the interaction types is constructed. This decision is supported by a simple network overlap analysis that measures the individual contribution of each type. This analysis is carried as follows. For each captured dataset, the intersection and the union sets of the edges between all pairs of interaction types are computed. Then, the overlap for each pair of interaction types is computed as $\mathrm{overlap}(A, B) = |A \cap B|/|A \cup B|$. The results reveal a fairly low overlap between every pair of types in all the datasets. The maximum overlap obtained is 11.8 %, between the *mentions* and *replies* interaction types in the Ireland2017 dataset and the minimum is 0.32 %, between *mentions* and *retweets* in the Pope2013-Spl dataset. This maximum is not surprising in the Ireland2017 dataset because of its long timespan of $\approx$245 days, therefore having a high chance of more overlapping interactions among users. In average, the overlap across all datasets is 2.39 % ($\sigma = 2.59$). This low overlap evidences that the multiple interaction types considered contribute independently to the overall network and hence can be combined.

## 3.5 GLOBAL PROPERTIES OF THE EXPERIMENTAL DATASETS

In this section, a general characterisation of the basic properties of the constructed ground-truth functional communities is presented with the purpose of providing a better understanding of the built ground-truth data for Twitter. The properties under study are descried below.

- **Community Size** is the number of users in the communities.
- **Membership Size** is the number of communities that a user belongs to.
- **Absolute Overlap Size** is the number of common users in a pair of overlapping communities.
- **Fraction of Overlap** is the fraction of the size of the overlap between any two communities over the size of the smaller of these communities.
- **Community Age** is the length of time between the first user interaction and the last.

The cumulative distribution function (CDF) of a real-valued random variable X is given by Equation 3.1 [Par18], where the right-hand side probability represents the chances that the random variable X is valued less than or equal to a value of interest x.

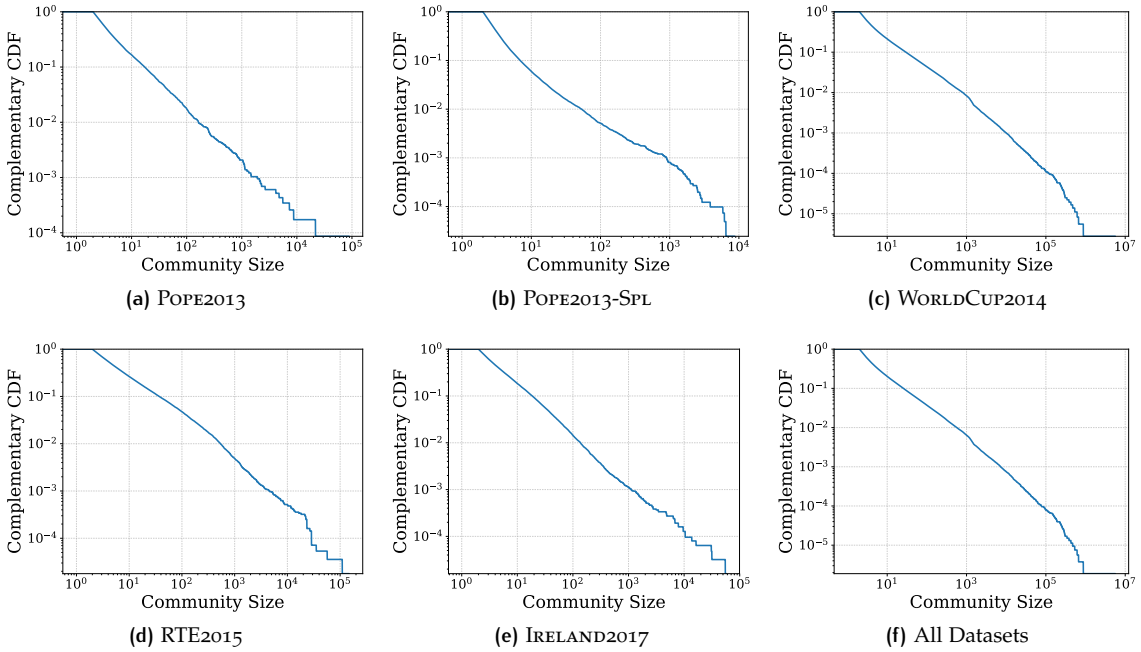
$$\mathrm{CDF}(X) = P(X \leqslant x) \tag{3.1}$$

However, to characterise the properties of the ground-truth datasets, it is more useful to study the opposite probability, i.e. how probable is that the random variable X is *above* a particular value x. This is known as the complementary cumulative distribution function (CCDF), and is defined in Equation 3.2. Most of the properties in this section are reported using the CCDF.

$$\mathrm{CCDF}(X) = P(X > x) = 1 - \mathrm{CDF}(X) \tag{3.2}$$

The first property under study is the **distribution of community sizes**, i.e. the number of users in the communities. The CCDF of the community sizes for each constructed dataset (considering all of the community types within) can be seen in Figure 3.4. Results considering all of the communities from all the datasets combined are also included.

It can be observed that all of the distributions are skewed with most ground-truth communities being small, e.g. sizes between 1 and 10. However larger communities also exist, e.g. sizes $\approx 10^4$ and $\approx 10^5$, and for the case of the WorldCup2014 dataset, up to $10^6$. These observations highlight the challenge of finding functional communities in Twitter.




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

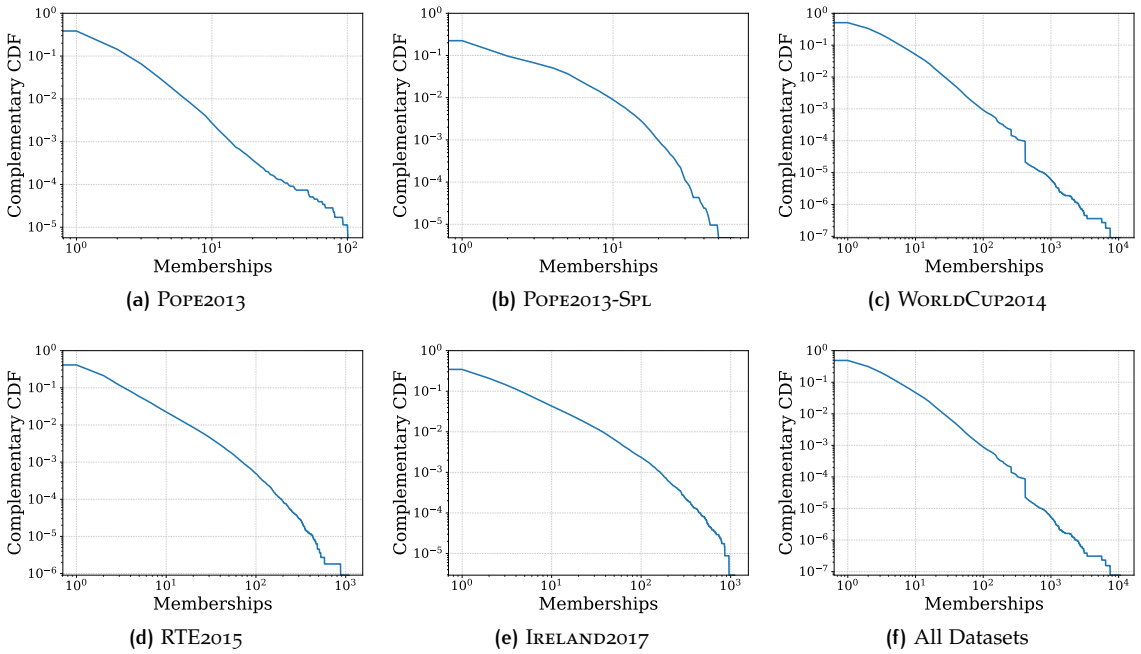
**Figure 3.4:** Complementary Cumulative Distribution Function (CCDF) using a log scale of the functional community sizes for each Twitter ground-truth dataset. A combined plot is also presented.

The complementary cumulative distribution function (CCDF) of the **user memberships sizes**, i.e. the number of communities that a user belongs to, can be seen in Figure 3.5, where the CCDF for each constructed dataset (considering all of the community types within) can be seen. Again, results considering all of the communities from all the datasets combined are also included.

In these results, user memberships of the functional communities are also very sparse, with many users belonging to just a handful of communities (<10) and few users belonging to many communities, e.g. generally $\approx 10^2$ and up to $\approx 10^4$ for the case of WORLDCUP2014.

Next the communities overlapping properties are investigated. For this, first the distribution of **absolute community overlap sizes**, i.e. the number of common users in a pair of overlapping communities, is presented using CCDF in Figure 3.6. Again, a skewed distribution following a power law is observed, similar to results described in [PDF+05] for *detected* communities in contrast to *ground-truth* data as it is investigated in this thesis.

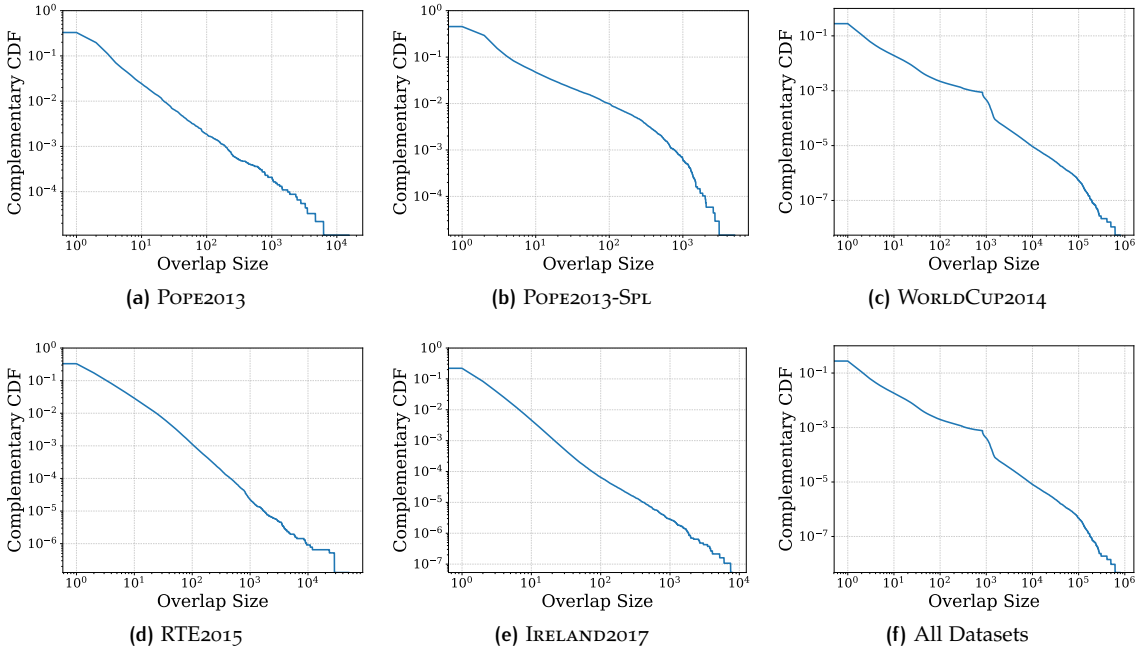
To complement these results, the **relative size of community overlaps** is also reported. This property is useful because it can characterise how the ground-truth functional communities actually overlap: in nested structures or only for a small number of users [YL15]. For this, the fraction of the size of the overlap between any two communities over the size of the smaller of these communities is measured using $f = |C_i \cap C_j| / min(|C_i|, |C_j|)$. If $f \approx 0$, then the majority of the communities do not overlap, and if $f \approx 1$, then the communities have a nested structure where the smaller communities are incorporated into the larger groups. In Figure 3.7, a his-




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

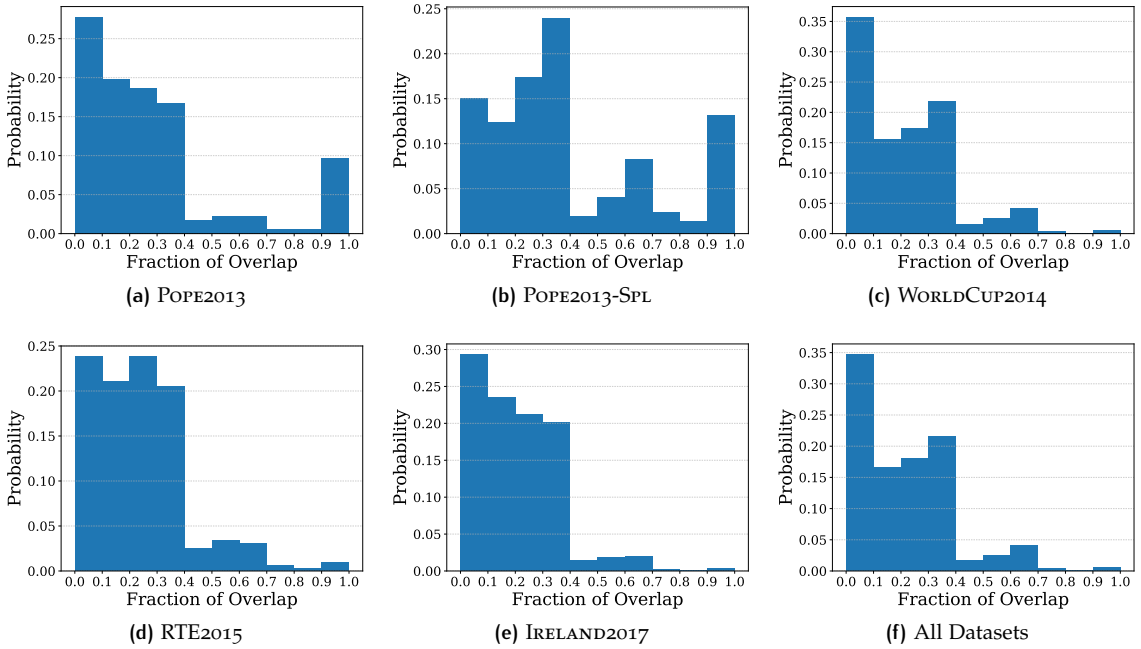
**Figure 3.5:** Complementary Cumulative Distribution Function (CCDF) using a log scale of the user membership sizes for each Twitter ground-truth dataset. A combined plot is also presented.




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

**Figure 3.6:** Complementary Cumulative Distribution Function (CCDF) of the functional community overlap sizes for each Twitter ground-truth dataset. A combined plot is also presented.




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

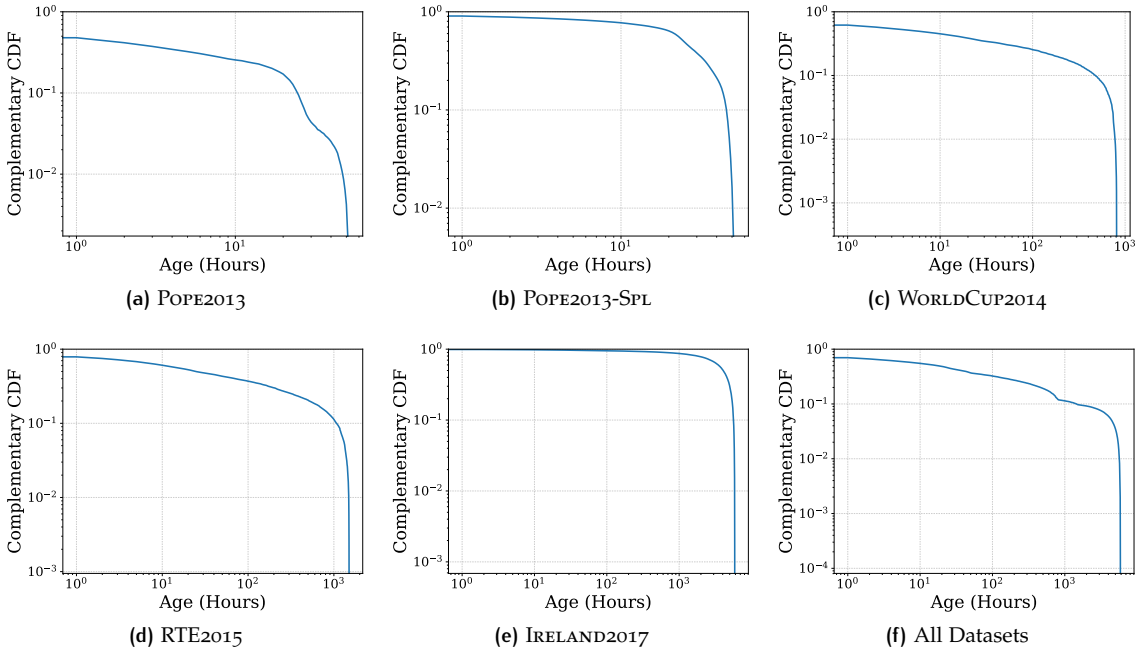
**Figure 3.7:** Histogram of the fraction of overlap for the ground-truth functional communities for each Twitter dataset. A combined plot is also presented.

togram of the fraction of overlap for each constructed dataset (considering all of the community types within) can be seen. Once more, results considering all of the communities from all the datasets combined are also included.

It can be observed in the histograms that many of the ground-truth functional communities indeed do not overlap in general. This is an expected result as many of these communities are built around very specific social objects, e.g. hashtags, mentioned users or specific locations. Nevertheless, a measurable number of ground-truth communities still exhibit some degree of nested overlapping $f \in [0.5, 0.7]$ for most of the cases and $f \geqslant 0.9$ for the POPE2013-SPL dataset, evidencing groups of users that participate using multiple social objects at the same time.

Finally, the complementary cumulative distribution function (CCDF) of the **functional community ages**, i.e. the length of time between the first user interaction and the last, can be seen in Figure 3.8 for each constructed dataset (considering all of the community types within). Results considering all of the communities from all the datasets combined are also included.

The reported community ages plots are heavily skewed and closely align to the timespan of each individual dataset, e.g. short ages for the POPE2013 and POPE2013-SPL datasets and longer ages for the IRELAND2017 dataset. This result suggests that there is a rich diversity of user activity in the functional communities across all the constructed datasets.




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

**Figure 3.8:** Complementary Cumulative Distribution Function (CCDF) of the functional community ages for each Twitter ground-truth dataset. A combined plot is also presented.

## 3.6 CHAPTER SUMMARY

In this chapter, the working definitions of the thesis for both, functional and structural user communities were introduced. Functional communities are defined as groups of users with a common and independent social function, e.g. fans of the same football team or people living in the same area, and structural communities are defined as groups of users that share a connectivity pattern connectivity in a network, e.g. their average node degree or clustering coefficient.

Because the social functions in Twitter are explicitly-labelled by the users themselves, functional communities are deemed as ground-truth community data in the thesis. Then, the task of evaluating community detection in microblogging is defined as evaluating the quality of the alignment of structural community definitions to the defined functional ground-truth.

Furthermore, a methodology for constructing ground-truth functional user communities from the Twitter microblogging platform based on eight types of social objects was also proposed. This methodology contemplates the building of both, functional communities and a user interactions network model based on four types of social objects for Twitter data. The user interactions network constitutes the working data for the structural community definitions in the thesis. Moreover, the contribution of each type of social function was measured and due to a low network overlap, ultimately a combined network approach is adopted.

Finally, the proposed approaches in this chapter were used to assemble a set of five real-world Twitter datasets. These datasets were captured considering different world-wide events during a wide range of years and all possible capturing methods from Twitter. The datasets were also characterised from a global perspective in terms of distribution of community size, membership size, overlap size and community ages, and form the basis for the experimental work in Chapter 4 and Chapter 5 of the thesis.

# 4 CHARACTERISING COMMUNITIES IN MICROBLOGGING

In this chapter, the **static scenario** of community detection in microblogging is investigated. The following main research question, proposed for this scenario in Chapter 1, is addressed.

(RQ2) → *How do existing structural community definitions accommodate to microblogging ground-truth communities, including their robustness to random perturbations?*

To provide an answer to this research question, the following research sub-questions are also proposed for this scenario, and are investigated in detail in this chapter.

- (RQ2.1) → *Do the defined ground-truth functional communities in (RQ1.1) evidence distinctive characteristics of structural communities, i.e. higher clustering coefficient, average degree, edge density and cohesiveness, in the associated networks of user interactions in (RQ1.2) in comparison to random groups with similar size and shortest-path distribution?*

- (RQ2.2) → *How well do state-of-the-art structural community definitions, e.g. based on triangle participation, conductance or modularity, align to the defined ground-truth functional communities in (RQ1.1) and (RQ1.2), including their robustness to random perturbations, e.g. member swapping, shrinking or expansion?*

First, the structural properties of the ground-truth functional communities defined in Chapter 3 are presented and discussed. Afterwards, an evaluation is carried in a static scenario of thirteen popular structural community definitions using the experimental Twitter datasets also introduced in Chapter 3. The goodness and robustness of the structural definitions for identifying the functional ground-truth under different perturbation strategies is also studied.

The identified contributions of this chapter are: (1) an in-depth characterisation, understanding and evaluation of structural properties for functional communities in microblogging social media, for the static scenario, and (2) a set of recommendations on community detection algorithms based on data-driven evaluation of Twitter user interactions networks.

The results show that community 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, other definitions such as Modularity are limited and do not perform well due to the sparsity and noisy characteristics of microblogging.

## 4.1 IDENTIFIABLE STRUCTURAL PATTERNS

Before attempting to perform community detection over user interactions networks from microblogging, a preliminary experiment to build evidence of identifiable structural patterns in these networks will be performed. The purpose of this preliminary analysis is to motivate the idea that microblogging user interactions networks do in fact have structural patterns that potentially can be identified using community detection approaches.

Therefore, to provide preliminary evidence of distinctive structural patterns in the network of user interactions, a comparative analysis of users in the ground-truth functional communities and randomly chosen connected nodes with the same path distribution is proposed [YL15]. If such distinctive connectivity patterns exist compared to randomly selected sets of connected nodes, then structural community detection algorithms likely will be able to discover the functional communities based on their network connectivity.

The sets of nodes to be used in this analysis for comparison will be now defined. For every ground-truth community $C_i$ (of any type) in the experimental datasets, a corresponding *non-community* $\tilde{C}_i$ is formed from the user interactions network with the following conditions:

1. $\tilde{C}_i$ must be of the same size than $C_i$
2. like every $C_i$, $\tilde{C}_i$ must also be connected
3. users in $\tilde{C}_i$ must have the same distribution of shortest path distances of $C_i$

In microblogging, the first and third constraints are not easily satisfiable. Both constraints are approached by first computing the $\chi^2$ distances [PW10] between the shortest path length histograms for every $C_i$ and for 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 $C_i$, the closest candidate $\tilde{C}_i$ that has at least 75 % of the size of $C_i$ is selected. Likewise, for the third constraint, if an exact match cannot be found, the closest candidate $\tilde{C}_i$ in descending order with the same distribution is selected. In case of multiple candidates, one is selected randomly.

After every ground-truth community $C_i$ is paired with a suitable non-community $\tilde{C}_i$, the set of structural properties that will be used to compare the structural patterns in the interactions network $G = (V, E, W)$ for both, ground-truth communities $C_i$ and non-communities $\tilde{C}_i$ are defined below. For this analysis, the edge weights $W$ will not be considered and remember that every community is guaranteed to have at least three members (refer to Chapter 3).

**CLUSTERING COEFFICIENT (CC)** [WS98]

This property is defined as the average local clustering coefficient of all the nodes in a community $C$ in the undirected sub-network $G_C = (V_C, E_C)$ induced by $C$. The local clustering coefficient for a node is the proportion of edges between the nodes within its neighbourhood divided by the number of edges that could possibly exist between them. In [WS98], this metric is used to measure how likely a set of nodes, i.e. a community $C$, is to form a *small-world* network, where the distances $L_i$ between two randomly chosen nodes follow the proportionality $L_i \propto \log(n)$ with $n$ the number of nodes in the network. A small-world network has relatively high clustering coefficient but small mean-shortest path length. The clustering coefficient is in the range $[0, 1]$, where values closer to one indicate a community with a stronger likelihood to being a clique, i.e. a complete graph.

**AVERAGE DEGREE (AVGDEG)** [Bon76]

This property measures the average degree of a set of nodes, i.e. a community $C$. The degree of a single node is the number of edges connected to that node and the average degree of a community $C$ is defined as $2|E_C|/|V_C|$, where $G_C = (V_C, E_C)$ is the undirected sub-network induced by the community $C$. It is in the range $[0, \infty)$, where higher is better.

**EDGE DENSITY** [RCC+04]

This property measures how similar a set of nodes, i.e. a community $C$, is to a clique structure. The edge density of a community $C$ is defined as $2|E_C|/(|V_C|(|V_C| - 1))$, where $G_C = (V_C, E_C)$ is the undirected sub-network induced by the community $C$. Density is in the range $[0, 1]$, where values closer to one are better.

**COHESIVENESS** [LLM10]

This property measures the fraction of total edges possible between a set of nodes, i.e. a community $C$, that are *non-bridging*. A non-bridge edge is such that when removed, the number of connected components in $C$ is preserved. This measure captures how resilient is the community $C$ and is in the range $[0, 1]$, where values closer to one indicate a stronger community that is harder to split or fragment.

**Table 4.1:** Ratio between structural properties of ground-truth functional communities and randomly chosen nodes with similar shortest path distribution for two representative experimental datasets.

**(a)** RTE2015

| C. Type | CC | AvgDeg | Density | Cohesiv | All $> 1.0$ |
|---|---|---|---|---|---|
| cities | 0.4034 | 1.0178 | 0.9169 | 0.5489 | |
| countries | 0.4365 | 0.9958 | 0.9510 | 0.6912 | |
| hashtags | 2.1117 | 1.2542 | 1.0885 | 2.0287 | Yes |
| mentions | 3.7619 | 1.7942 | 1.3538 | 3.1683 | Yes |
| places | 0.3981 | 0.9914 | 0.9329 | 0.4795 | |
| quotes | 2.3291 | 1.3907 | 1.1491 | 2.2839 | Yes |
| retweets | 2.8460 | 1.6003 | 1.1834 | 2.6283 | Yes |
| urls | 2.6746 | 1.2983 | 1.1495 | 2.4909 | Yes |
| **Average** | **1.8702** | **1.2928** | **1.0906** | **1.7900** | **Yes** |

**(b)** IRELAND2017

| C. Type | CC | AvgDeg | Density | Cohesiv | All $> 1.0$ |
|---|---|---|---|---|---|
| cities | 22.2567 | 1.1828 | 1.0691 | 12.4548 | Yes |
| countries | 8.6811 | 1.0263 | 1.0205 | 6.8682 | Yes |
| hashtags | 31.8735 | 1.2197 | 1.1139 | 15.7759 | Yes |
| mentions | 48.7442 | 1.4785 | 1.2394 | 22.7188 | Yes |
| places | 11.1738 | 1.0794 | 1.0380 | 7.5388 | Yes |
| quotes | 38.2470 | 1.3039 | 1.1757 | 20.3443 | Yes |
| **Average** | **26.8294** | **1.2151** | **1.1094** | **14.2835** | **Yes** |

Each structural property p above is computed for every $C_i$ and $\tilde{C}_i$, and then the average ratio $r = p(C_i)/p(\tilde{C}_i)$ is computed for all community types in each experimental dataset. If $r > 1.0$, then a measurable difference in the structural property p for $C_i$ compared to $\tilde{C}_i$ can be asserted.

The results for two representative datasets, RTE2015 and IRELAND2017, are in Table 4.1[1]. The WORLDCUP2014 dataset has similar results as RTE2015. In all datasets, the property ratios averaged over all communities is larger than one. Furthermore, IRELAND2017 is the only dataset where all the community types have a property ratio greater than one. In contrast, the remaining datasets do not exhibit a distinguishable ratio for the *countries*, *cities* and *places* community types. This is explained by the low number of communities built for these types (refer to Table 3.3).

In the example of RTE2015, ground-truth functional communities have, in average, 87 % higher clustering coefficient, 29 % higher average degree, 9 % higher edge density and 79 % higher cohesiveness than their respective non-communities. In the case of IRELAND2017, the ground-truth has, in average, ≈27 times better CC, 22 % higher average degree, 11 % higher edge density and ≈14 times better cohesiveness. All the obtained results suggest that the ground-truth functional communities have more community-like structural properties compared to randomly chosen nodes in the same network with nearly the same shortest paths distribution.

1 All the structural properties can be found in Section B.1 in the appendices.

In general, the results show that the ratio for each defined property are $r \geqslant 1.0$ in the majority of the ground-truth community types and datasets. The *mentions* community type excels in every dataset, where a high ratio between communities $C_i$ and non-communities $\tilde{C}_i$ can be observed, suggesting that functional communities with a third person as functional object are easier to discover than other types of social objects. Similar is the *hashtags* type, where it is only weak ($r < 1.0$) in the POPE2013 dataset. This is unexpected because it suggests that users do not form strong communities around hashtags, despite the Pope Conclave event being highly susceptible to discussions around specific topics. The IRELAND2017 dataset has very strong differentiable hashtags communities, e.g. $r > 30$ for CC and $r > 15$ for Cohesiveness. This observation is explained by more persistent interactions ($A_i$) in this dataset as shown in Table 3.1.

The *retweets* and *urls* functional types exhibit structurally distinguishable functional communities ($r > 1.0$) in the majority of the experimental datasets and consistent with [KLP+10], where retweeting and media links are regarded as core activities for news diffusion in Twitter. The *countries*, *cities* and *places* functional types in general contain few distinguishable communities, with the exception of IRELAND2017. This can be explained by the low signal of Tweets that actually contain useful location information [HHS+11]. The IRELAND2017 dataset is captured using the location filter, therefore every Tweet in this dataset is embedded with location data. Nevertheless, the *countries* type was found to be distinctive enough (in most cases $r > 1.0$), suggesting that this abstraction is the most suitable for building functional communities based on location.

Finally, the *quotes* type is a recent functionality in Twitter, therefore is not present in datasets captured before 2015. Nonetheless, for the datasets that contain functional communities based on quotes, there is strong measurable structure difference ($r > 2.0$ for CC and Cohesiveness), suggesting that Twitter users seem to interact closely around quotes of interest to them.

## 4.2 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 a desired structure. In this thesis, thirteen commonly used community scoring functions pre-classified into four families are considered for evaluation.

In general, for a set of nodes $C$, a scoring function $f(C)$ measures the quality of $C$ as a *structural community* in an undirected network $G = (V, E)$. Many – but not all – community detection algorithms do not take any edge weights into consideration for the discovery of communities.

Let $n_c$ and $m_c$ be the number of nodes and edges respectively 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, the scoring functions under evaluation in the thesis are introduced below.

#### BASED ONLY ON INTERNAL CONNECTIVITY

This family of scoring functions measure community structures only based on their internal connectivity, i.e. they only consider the nodes in the community and their edges, without considering any part of the rest of the network.

- **Density** [RCC+04] is the fraction of total edges possible in C that are actually present.

  $f(C) = density(C) = 2m_c/(n_c(n_c - 1))$

  This score is in the range $[0, 1]$ and a higher value means better communities.

- **Edges Inside** [RCC+04] is simply the number of edges present in C.

  $f(C) = edgesInside(C) = m_c$

  This score is in the range $[0, \infty)$ and a higher value means better communities.

- **Average Degree** [RCC+04] is the average degree of the nodes in C.

  $f(C) = avgDeg(C) = 2m_c/n_c$

  This score is in the range $[0, \infty)$ and a higher value means better communities.

- **Fraction over Median Degree (FOMD)** [YL15] is the fraction of nodes in C that have degree higher than $d_m$, where $d_m$ is the median degree of all nodes $v \in V$ of the network. Please note that $d_m$ is an *external* factor to C, however it is also constant for all communities, therefore FOMD is still considered an internal connectivity score.

  $f(C) = fomd(C) = |\{u : u \in C, |\{(u, v) : v \in C\}| > d_m\}|/n_c$

  This score is in the range $[0, 1]$ and a higher value means better communities.

- **Triangle Participation Ratio (TPR)** [YL15] is the fraction of nodes in C that belong to a triad. A triad is a set of three nodes that are fully connected to each other in C.

  $f(C) = tpr(C) = |\{u : u \in C, u \text{ is in a triad}\}|/n_c$

  This score is in the range $[0, 1]$ and a higher value means better communities.

#### BASED ONLY ON EXTERNAL CONNECTIVITY

This family of scoring functions measure community structures only based on their external connectivity, i.e. they only consider the nodes and edges at the boundaries of the community, without considering the rest of the community.

- **Expansion** [RCC+04] quantifies the number of edges per node on the boundary of $C$.

  $f(C) = expansion(C) = b_c/n_c$

  This score is in the range $[0, \infty)$ and a lower value means better communities.

- **Cut Ratio** [For10] is the fraction of existing edges (out of all possible edges) leaving $C$.

  $f(C) = cutRatio(C) = b_c/(n_c(n - n_c))$

  This score is in the range $[0, 1]$ and a lower value means better communities.

#### BASED ON INTERNAL AND EXTERNAL CONNECTIVITY

This family of scoring functions measure community structures based on both, their external and internal connectivity, i.e. they consider the nodes and edges that are inside and outside of the community simultaneously.

- **Conductance** [SM00] is the fraction of total edge volume that is on the boundary of $C$.

  $f(C) = conductance(C) = b_c/(2m_c + b_c)$

  This score is in the range $[0, 1]$ and a lower value means better communities.

- **Normalized Cut** [SM00] quantifies the effort of cutting edges at the boundary of the community $C$ as the fraction of the total edges at the boundary to all the nodes in the network. The normalized cut is an unbiased score with respect to isolated nodes.

  $f(C) = normCut(C) = b_c/(2m_c + b_c) + b_c/(2(m - m_c) + b_c)$

  This score is in the range $[0, 1]$ and a lower value means better communities.

- **Maximum Out Degree Fraction (Max ODF)** [FLG00] is the maximum fraction of edges in the community $C$ that point outside of the community, i.e. its boundary.

  $f(C) = maxOdf(C) = \max_{u \in C}[|\{(u, v) \in E : v \notin C\}|/d(u)]$

  This score is in the range $[0, 1]$ and a lower value means better communities.

- **Average Out Degree Fraction (Avg ODF)** [FLG00] is the average fraction of edges in the community $C$ that point outside of the community, i.e. its boundary.

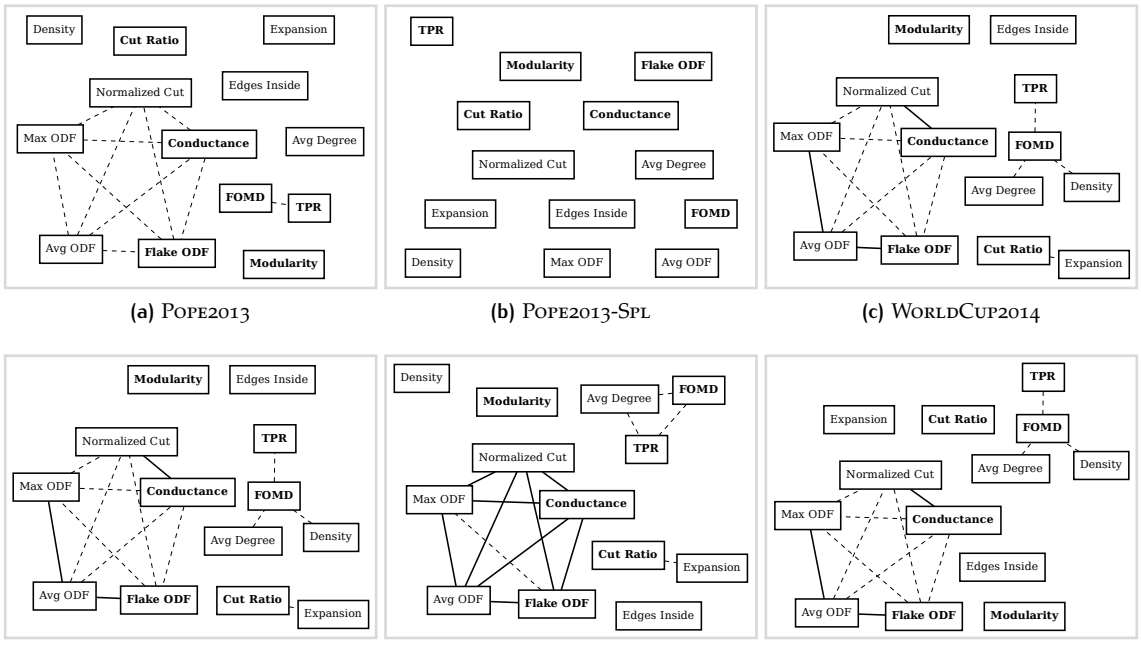
  $f(C) = avgOdf(C) = \sum_{u \in C}[|\{(u, v) \in E : v \notin C\}|/d(u)]/n_c$

  This score is in the range $[0, 1]$ and a lower value means better communities.

- **Flake Out Degree Fraction (Flake ODF)** [FLG00] is the fraction of nodes in the community $C$ that have fewer edges pointing inside than outside of the community.

  $f(C) = flakeOdf(C) = |\{u : u \in C, |\{(u, v) \in E : v \in C\}| < d(u)/2\}|/n_c$

  This score is in the range $[0, 1]$ and a lower value means better communities.




**(a)** POPE2013 **(b)** POPE2013-SPL **(c)** WORLDCUP2014

**(d)** RTE2015 **(e)** IRELAND2017 **(f)** All Datasets

**Figure 4.1:** Scoring functions clustered by correlation for each Twitter ground-truth dataset. Weak links ($\rho \geqslant 0.3$) are dashed and strong links ($\rho \geqslant 0.6$) are solid. A combined plot is also presented.

#### BASED ON NETWORK MODEL

This family of scoring functions measure community structures based on their similarity to a defined network model that mimics the desired structural characteristics.

- **Modularity** [New06a] is the difference between the number of edges $m_c$ and those expected $E[m_c]$ in a random graph with identical degree sequence, i.e. a null model.

  $f(C) = modularity(C) = m_c - E[m_c]$

  This score is in the range $[-0.5, 1)$ and a higher value means better communities.

The relationship between these scoring functions in the experimental Twitter ground-truth data is now initially explored. To investigate the individual contribution of these scoring functions, first each score $f(C)$ for each of the constructed ground-truth functional communities $C$ is computed. Then, a correlation matrix based on the Pearson coefficient is constructed and filtered to unveil the correlation at different degrees between the scoring functions. Following the guidelines in [Eva96], $\rho \geqslant 0.3$ and $\rho \geqslant 0.6$ are adopted as thresholds for weak and strong correlation respectively. The results for each experimental dataset and all the ground-truth data combined as well can be seen in Figure 4.1. In the graphs, weak correlation between scores is represented using dashed links and strong correlation with solid connections. All the Pearson coefficients computed for the matrices were found as significant with a small p-value $\leqslant 0.05$.

With the sole exception of the POPE2013-SPL dataset, which contains many very small communities due to its random sampling nature, all the scores in the mixed – i.e. internal and external – connectivity family correlate together at different degrees. The internal-only connectivity scores remained more fragmented, with FOMD and TPR having a stronger correlation. Density and Average Degree still correlate to the above, however in a lesser degree. The Edges Inside score remained isolated, even from its close relative Average Degree. This suggests that, for the case of Twitter, simple edge-based scores perform very poorly in presence of highly sparse communities. For the case of the external-only connectivity scores, they correlate in the larger WORLDCUP2014, RTE2015 and IRELAND2017 datasets, but not on the Pope Event datasets. This is also attributed to the ground-truth communities in these datasets not persisting long enough in time.

In general, with few exceptions, the majority of the scores correlate within their own predefined families, revealing roughly four principal correlated groups. Therefore, this thesis focus on six representative scoring functions (in bold in Figure 4.1) derived from these correlated groups: (1) FOMD, (2) TPR, (3) Cut Ratio, (4) Conductance, (5) Flake ODF, and (6) Modularity.

The above representative scoring functions were selected by considering how recent in the literature and elaborated in their definition they are. For instance, FOMD and TPR from the internal connectivity family are the most recently proposed [YL15] and also less simplistic than Density, Edges Inside and Average Degree. Likewise, Cut Ratio (external connectivity family) was proposed more recently [For10] and captures more information than Expansion. In the mixed connectivity family, Conductance (preferred) and Normalized Cut were compared and found similar by J. Shi and Malik [SMoo], while Max ODF, Avg ODF and Flake ODF (preferred) were evaluated and proposed by Flake, Lawrence, and Giles [FLGoo].

This exploratory experiment suggests that despite having numerous scoring functions to measure structural communities, they still correlate in the Twitter example of microblogging.

## 4.3 GOODNESS OF COMMUNITY DETECTION

The representative community scoring functions introduced before will be now evaluated in terms of their quality to discover ground-truth functional communities in microblogging streams, represented by Twitter. In this experiment, goodness metrics will be used that capture the notion that good communities should be compact, well connected and well isolated from the rest of the network. The difference between these 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 or community. A group with high goodness does not imply a good scoring function value but a good community score should have a high goodness metric.

Four goodness metrics $g(C)$ are considered. Three of them (Clustering Coefficient, Density and Cohesiveness) were previously introduced as structural properties in Section 4.1. A fourth additional goodness metric is introduced below to complement these structural properties.

**SEPARABILITY** [For10]

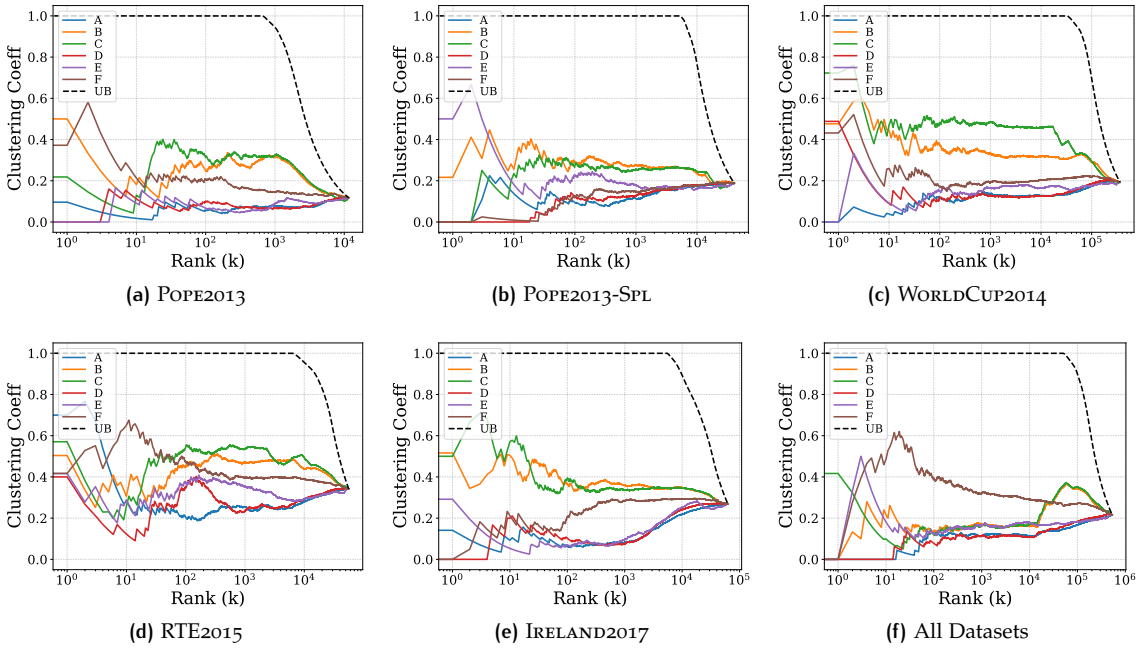
> This goodness metric measures how well-distanced is a set of nodes, i.e. a community C, in terms of the ratio between the internal and external edges between them. This measure captures the intuition that good communities should be well-distanced from each other. Using the same notation as for the scoring functions, the separability of a community C is defined as $|E_C|/b_c$, where $G_C = (V_C, E_C)$ is the undirected sub-network induced by the community C. Separability is in the range $[0, \infty)$, where higher values are better.

To evaluate the goodness of the scoring functions for the Twitter ground-truth data, the experiment is proposed as follows. For each dataset and community type, the ground-truth functional communities $C_i$ are ranked using the six selected scoring functions $f(C_i)$ in descending order. Then, the cumulative moving average (CMA) of each goodness metric $g(C_i)$ is observed 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. Conversely, a poor community scoring function would produce a k-dependent constant CMA.

Figures 4.2, 4.3, 4.4 and 4.5 respectively show the results for the ranked Clustering Coefficient, Density, Cohesiveness and Separability for each ground-truth dataset, including a plot with all the data combined. The upper bound curve, i.e. the CMA of a goodness metric ranked by the same goodness metric, which represents a perfect ranking is also provided for reference.

### 4.3.1 Clustering Coefficient

For the **Clustering Coefficient** goodness metric in Figure 4.2, the FOMD (B) and TPR (C) scores prevail with the better performance, except in the case when all the data is combined. Because the combined plot represents an overall average of every ranking, the Modularity (F) score stands artificially higher due to its performance in RTE2015. Otherwise, Modularity (F) exhibits a poor ranking and even constant in some cases, suggesting that this score is rather unstable and does not prefer nor reject dense communities in Twitter. On the other hand, Cut Ratio (A),




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

**Figure 4.2:** Ranked Clustering Coefficient by CMA for each Twitter ground-truth dataset. A combined plot is also presented. Scores: Cut Ratio (A), FOMD (B), TPR (C), Conductance (D), Flake ODF (E), Modularity (F) and their Upper Bound (UB).

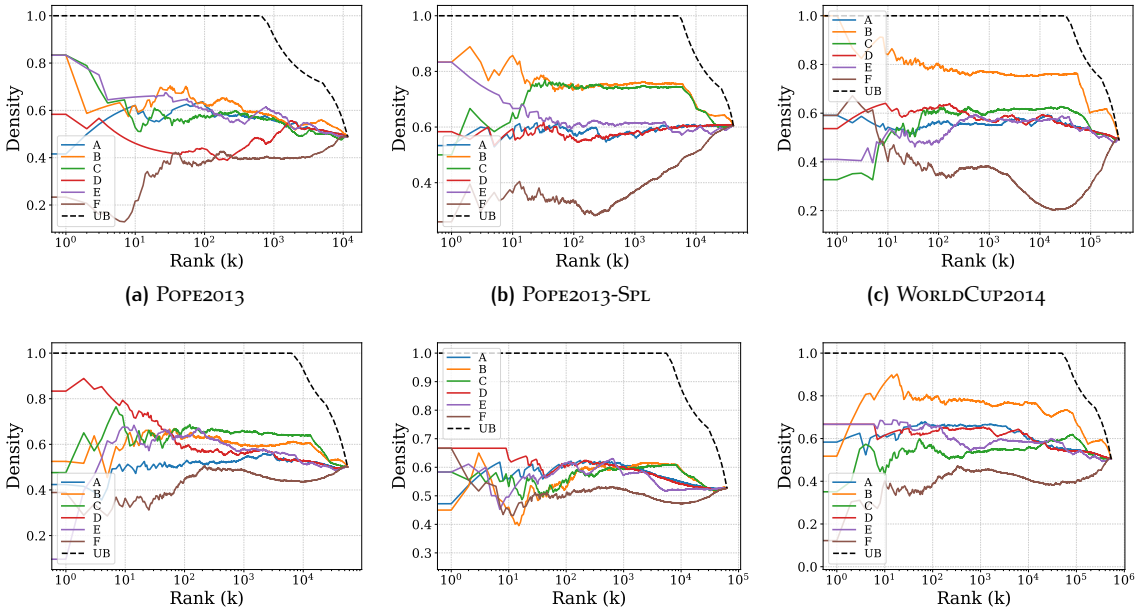
Conductance (D) and Flake ODF (E) show instead an inverse ranking, suggesting that these scores tend to reject ground-truth communities that are more tightly connected.

### 4.3.2 Density

For the **Density** goodness metric in Figure 4.3, the results are similar to Clustering Coefficient. The FOMD (B) and TPR (C) scores prevail with the better performance, even in the combined data case. Modularity (F) strongly disagrees with the goodness metrics regarding dense communities in the Twitter ground-truth, which is a manifestation of the well-known resolution limit of the Modularity score [FB07]. Cut Ratio (A), Conductance (D) and Flake ODF (E) all exhibit near constant ranking, evidencing that they do not prefer denser communities in Twitter.

### 4.3.3 Cohesiveness

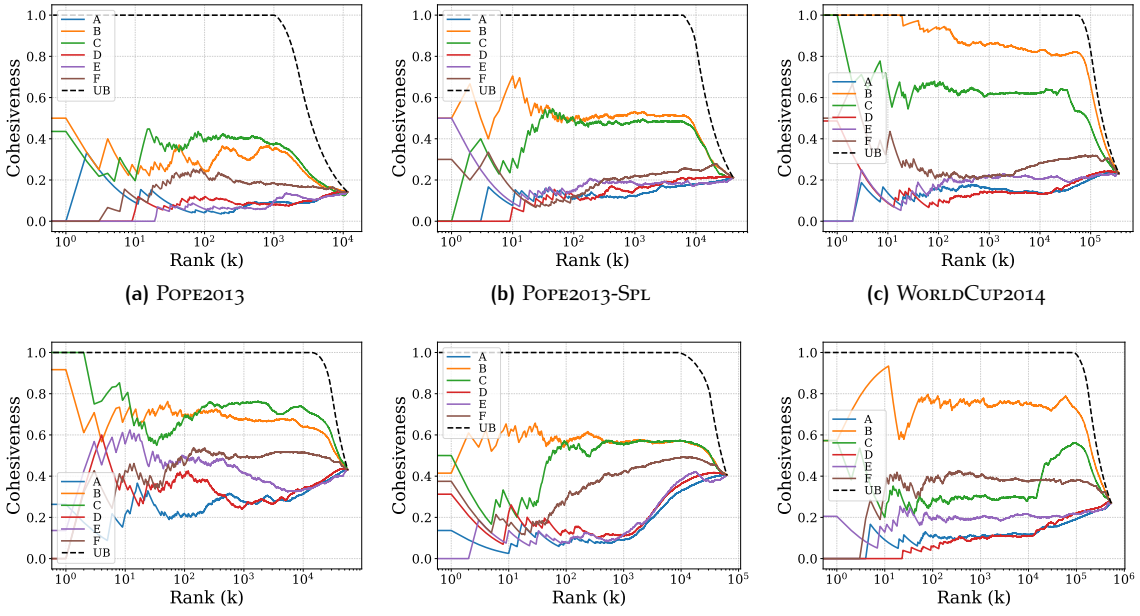
For the **Cohesiveness** goodness metric in Figure 4.4, the results are again similar to Clustering Coefficient and Density. The FOMD (B) and TPR (C) scores are able to easily outperform the other scores in all the datasets, specially evident for WORLDCUP2014. For the overall case of all




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

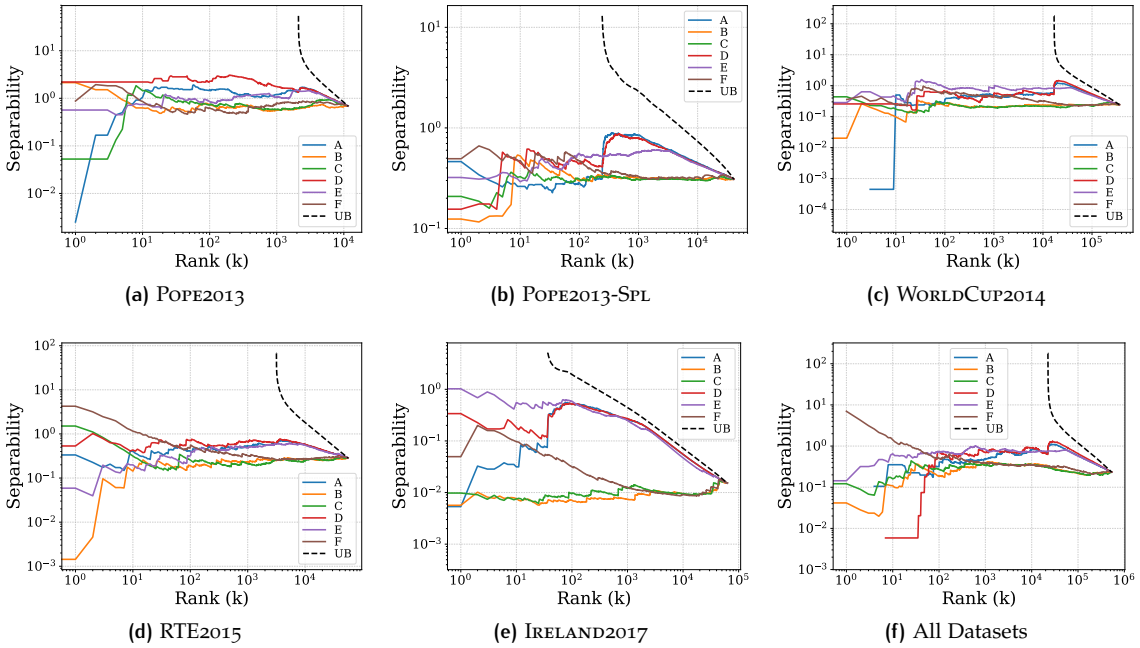
**Figure 4.3:** Ranked Density by CMA for each Twitter ground-truth dataset. A combined plot is also presented. Scores: Cut Ratio (A), FOMD (B), TPR (C), Conductance (D), Flake ODF (E), Modularity (F) and their Upper Bound (UB).




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

**Figure 4.4:** Ranked Cohesiveness by CMA for each Twitter ground-truth dataset. A combined plot is also presented. Scores: Cut Ratio (A), FOMD (B), TPR (C), Conductance (D), Flake ODF (E), Modularity (F) and their Upper Bound (UB).




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

**Figure 4.5:** Ranked Separability by CMA for each Twitter ground-truth dataset. A combined plot is also presented. Scores: Cut Ratio (A), FOMD (B), TPR (C), Conductance (D), Flake ODF (E), Modularity (F) and their Upper Bound (UB).

the data considered, Modularity (F) has reasonable good performance, surpassing TPR (C) in most of the ranking. The group in the external and mixed connectivity families, Cut Ratio (A), Conductance (D) and Flake ODF (E) exhibit inverse ordering in the ranking. This observation suggests that these scores prefer sparse communities, revealing their inability to properly capture cohesive groups in the Twitter microblogging scenario.

### 4.3.4 Separability

Lastly, for the **Separability** goodness metric in Figure 4.5, the community scores show an opposite behaviour in comparison to the previous goodness metrics. Cut Ratio (A), Conductance (D) and Flake ODF (E) have a close fit to the upper bound in Separability, specially for the IRELAND2017 dataset, while FOMD (B) and TPR (C) show instead an inverse ordering, suggesting that these two prefer more dense communities. If the analyst desires denser, more packed communities regardless of separation in the static scenario, FOMD and TPR should be preferred.

Table 4.2: Aggregated scoring ranking by goodness metrics using the Borda voting method for all ground-truth datasets. Best ranked scoring functions for each goodness metric are in bold.

| Family | Score | CC | Cohesiveness | Density | Separability |
|---|---|---|---|---|---|
| External | Cut Ratio | 5.4589 | 5.7208 | 3.3097 | 2.2203 |
| Internal | FOMD | 2.0974 | **1.0014** | **1.0001** | 5.0531 |
| Internal | TPR | **1.1422** | 2.2890 | 3.4189 | 5.6131 |
| Mixed | Conductance | 4.1373 | 3.9755 | 2.8899 | **1.3169** |
| Mixed | Flake ODF | 5.2929 | 5.2416 | 4.8373 | 2.4659 |
| Net-Model | Modularity | 2.8712 | 3.0755 | 5.5441 | 4.3307 |

### 4.3.5 Goodness Metrics Ranking

To complement the goodness analysis, the ability of the scoring functions to rank the ground-truth communities based on the goodness metrics is also investigated. For each goodness metric $g(C)$ and scoring function $f(C)$ in all of the ground-truth datasets, the ranking position of each score is observed in comparison to every other scoring function at every rank $k$. For example, in Figure 4.2 for clustering coefficient at $k = 10^2$, the scores are ranked as: $1^{st}$ TPR (C), $2^{nd}$ FOMD (B), $3^{rd}$ Modularity (F), $4^{th}$ Conductance (D), $5^{th}$ Flake ODF (E) and $6^{th}$ Cut Ratio (A). Therefore, for every $k$, the six scores are ranked and aggregated using the Borda voting method [Saa12] to obtain an unified ranking that quantifies the ability of each scoring function to find *good communities*. The averaged results for all the ground-truth datasets are in Table 4.2, where ranks $\approx 1.0$ (in bold in the table) indicate scoring functions adequate for each goodness criteria.

The scoring functions based on internal structural information, i.e. from the internal and mixed families, demonstrated to be the best performing in this experiment. Overall, to identify more clustered, dense and cohesive communities in Twitter, FOMD and TPR are the better choices for structural scoring functions. If dense but more separated communities are desired by the analyst, then Conductance or Cut Ratio should be considered.

## 4.4 ROBUSTNESS OF COMMUNITY DETECTION

Good scoring functions should be stable under small perturbations and reduce their performance under strong disturbance. To further study the goodness of the structural scoring functions for Twitter in the static scenario, their robustness and sensitivity in presence of different random perturbations to the ground-truth functional communities is now investigated. The following random perturbation strategies for communities [YL15] are adopted and evaluated.

#### NODE SWAP

This perturbation strategy simulates the effect of members diffusing from the community through the network. First, a random edge $(u_i, u_j), u_i \in C, u_j \notin C$ is chosen, then the nodes $u_i$ and $u_j$ are swapped. This causes $u_i$ to abandon $C$ and $u_j$ to join it. Node Swap does not change the size of the community, but it may internally fragment it.

#### RANDOM

This perturbation strategy internally perturbs communities by swapping a random member $u_i \in C$ with a random non-member $u_j \notin C$. In contrast to Node Swap, the nodes being swapped do not need to be directly connected. Similar to Node Swap, this strategy does not change the size of the community but it may internally fragment it.

#### EXPAND

This perturbation strategy increases the size of communities by choosing random non-members $u_j \notin C$ that are connected to members $u_i \in C$, and incorporating them into $C$. This action decreases the quality of the community but it preserves its connectedness.

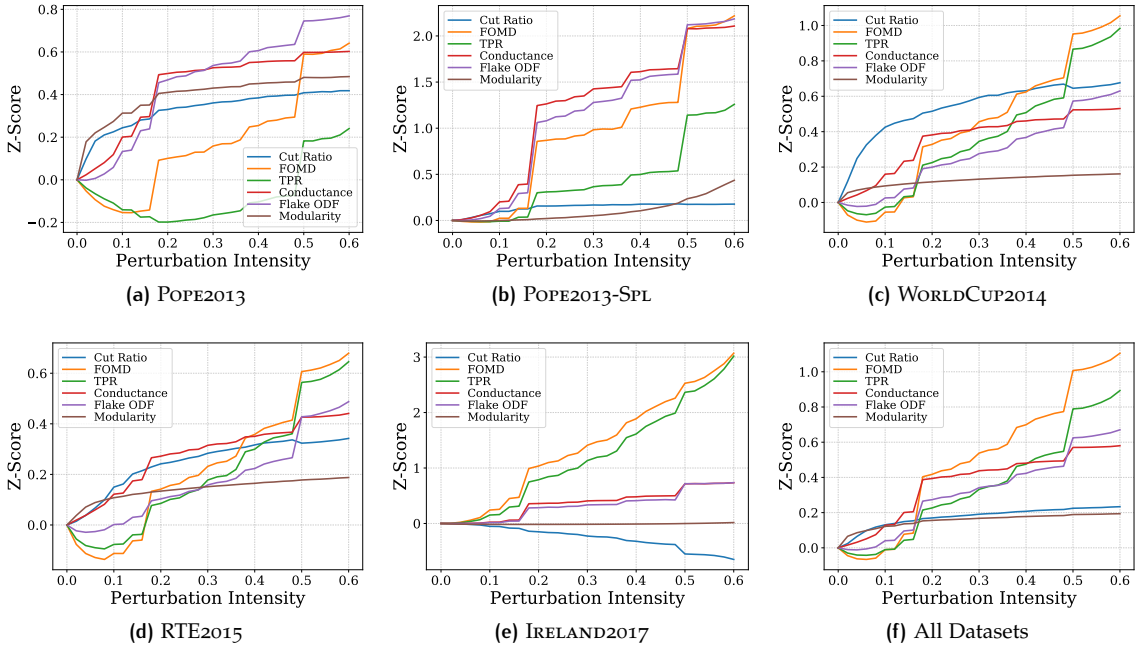
#### SHRINK

This perturbation strategy decreases the size of communities by choosing random boundary edges $(u_i, u_j), u_i \in C, u_j \notin C$ and removing the user $u_i$ from $C$. Similar to Expand, this perturbation preserves the connectedness of the community.

The above strategies can be controlled using an intensity parameter $p$, that specifies the number of times, i.e. $p|C|$, the given perturbation is applied to a community $C$. For example, for the Node Swap strategy $p = 0.60$ means exchanging 60 % of the members of $C$.

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, to measure the difference of a score $f$ applied to $C$, i.e. $f(C)$, and the same score applied to the perturbed version of $C$, i.e. $f(h(C, p))$, the Z-score statistic (units of standard deviation) is adopted in Equation 4.1, where $E[\cdot]$ is the expectancy operator (the mean) and $Var[\cdot]$ is the variance operator, both applied over all the ground-truth communities $C_i$. Large Z-scores indicate perturbed scoring functions $f(h(C, p))$ more separated from the unperturbed versions $f(C)$, therefore the Z-scores are expected to increase along the perturbation intensity $p$.

$$Z(f, h, p) = \frac{E[f(C_i) - f(h(C_i, p))]}{\sqrt{Var[f(h(C_i, p))]}} \tag{4.1}$$




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

**Figure 4.6:** Z-scores of intensities for the Node Swap perturbation strategy applied to all community types for each Twitter ground-truth dataset. A combined plot is also presented.

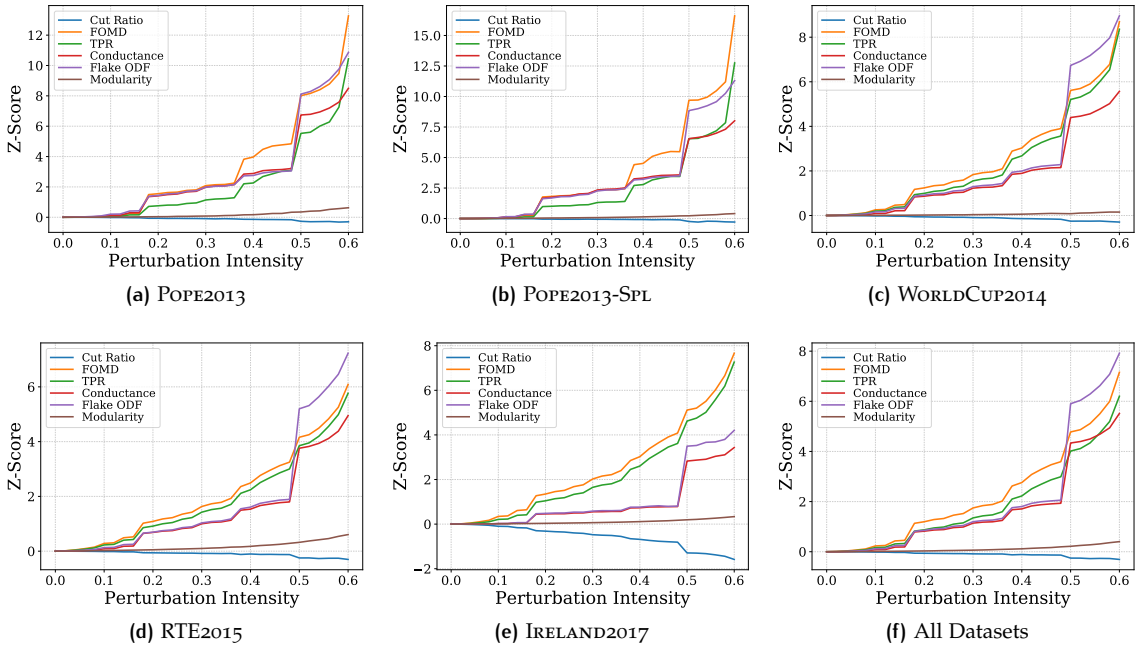
The TPR, FOMD and Modularity scores need to be inverted, i.e. change sign, to ensure that all the scores under evaluation have the same interpretation, i.e. a higher score value is considered better. Furthermore, due to the random nature of the perturbations strategies, the experiment is repeated 20 times and the resulting Z-scores averaged.

With all the above, the perturbation experiment is now defined as follows. The perturbation intensity is varied in the range $p \in [0.01, 0.60]$, e.g. in the Node Swap strategy this means exchanging between 1 and 60 % of the members of a community, and observe the averaged Z-score across all ground-truth functional communities in all community type and datasets.

Figures 4.6, 4.7, 4.8 and 4.9 respectively show the averaged Z-score results for the Node Swap, Random, Shrink and Expand perturbation strategies under the proposed intensities for each ground-truth dataset, including a plot with all the data combined.

### 4.4.1 Node Swap

For the **Node Swap** perturbation in Figure 4.6, the TPR and FOMD scores perform the best in all the datasets with the longer timespan (WORLDCUP2014, RTE2015 and IRELAND2017), followed by the Conductance and Flake ODF scores. In the case of the Pope Event datasets, Conductance and Flake ODF instead are observed as more robust scores, i.e. they degrade more gracefully in




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

**Figure 4.7:** Z-scores of intensities for the Random perturbation strategy applied to all community types for each Twitter ground-truth dataset. A combined plot is also presented.

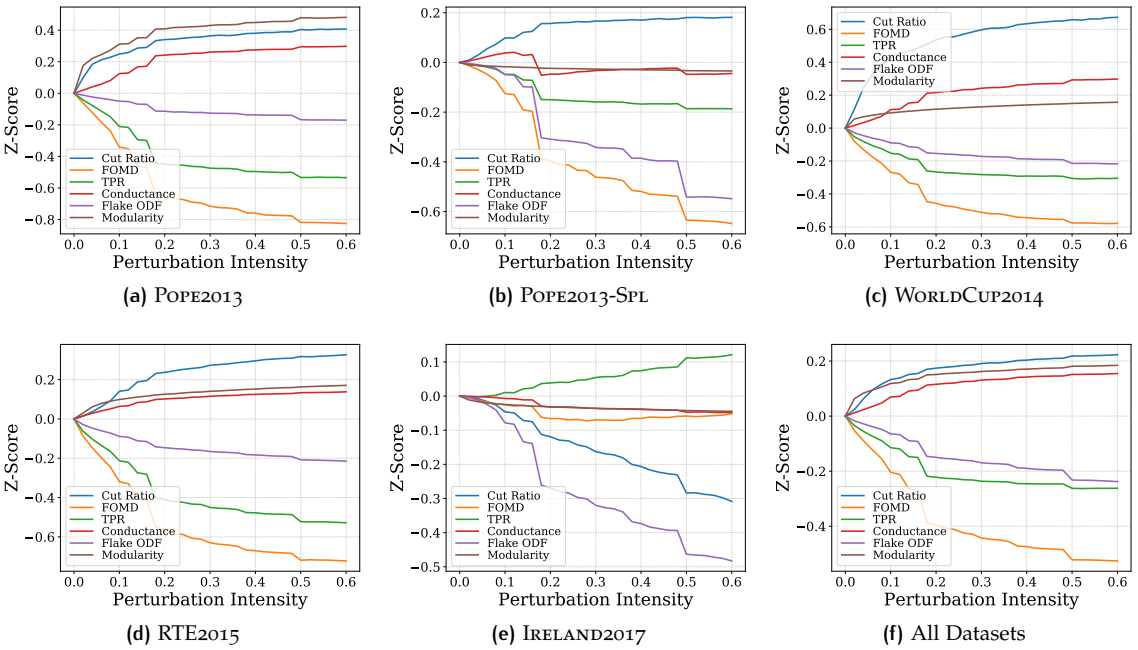
presence of stronger node swapping perturbation. In contrast, Modularity and Cut Ratio do not degrade as gracefully – particularly Modularity – when the perturbation is increased, revealing their inability to handle noisy data in Twitter.

### 4.4.2 Random

For the **Random** perturbation in Figure 4.7, the internal and mixed connectivity families of scores – FOMD, TPR, Conductance and Flake ODF – consistently perform the best, with the internal family being ultimately the most robust (for example in the IRELAND2017 dataset). Cut Ratio and Modularity perform the worst in presence of strong noise, with their Z-scores having very small variation under higher levels of perturbation.

### 4.4.3 Expand and Shrink

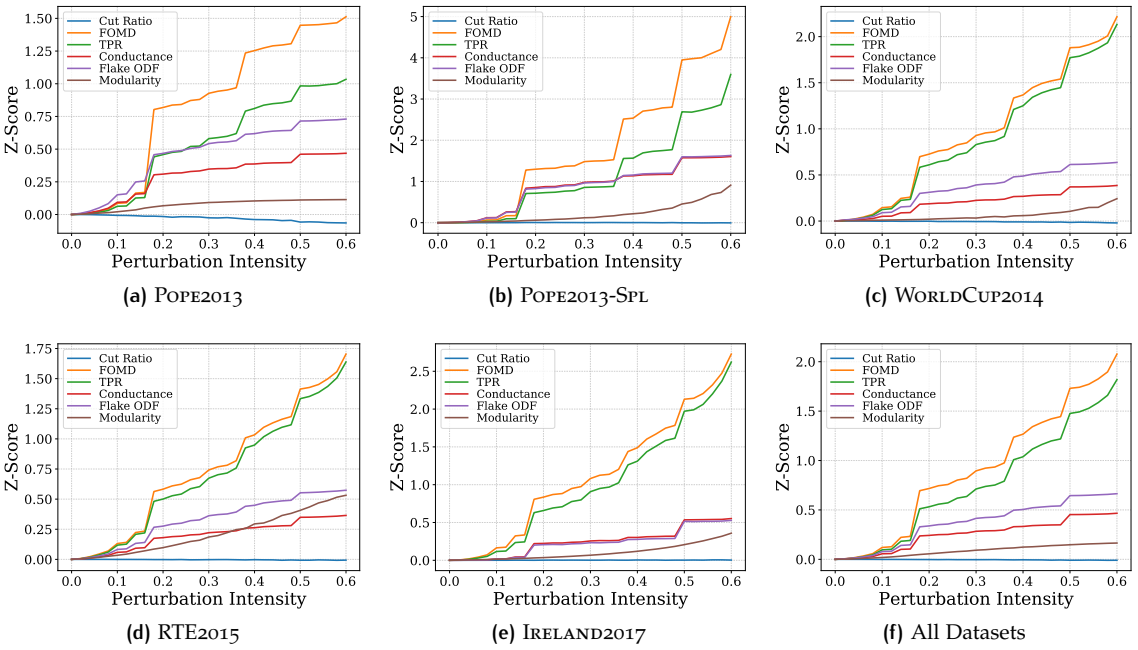
Lastly, the **Expand** and **Shrink** perturbations results seen in Figure 4.8 and Figure 4.9 also reveal TPR and FOMD as generally robust scores for Twitter functional communities, specially the Shrink perturbation. The Cut Ratio score is unable to handle communities that get smaller in Twitter, however its robustness improves when communities expand. Modularity has consistent




(a) Pope2013 (b) Pope2013-Spl (c) WorldCup2014

(d) RTE2015 (e) Ireland2017 (f) All Datasets

**Figure 4.8:** Z-scores of intensities for the Expand perturbation strategy applied to all community types for each Twitter ground-truth dataset. A combined plot is also presented.




(a) Pope2013 (b) Pope2013-Spl (c) WorldCup2014

(d) RTE2015 (e) Ireland2017 (f) All Datasets

**Figure 4.9:** Z-scores of intensities for the Shrink perturbation strategy applied to all community types for each Twitter ground-truth dataset. A combined plot is also presented.

**Table 4.3:** Average absolute increment of Z-score between small ($p = 0.04$) and large ($p = 0.20$) community perturbations. Largest differences (most robust and sensitive scores) are in bold.

| Family | Score | N.Swap | Random | Expand | Shrink |
|---|---|---|---|---|---|
| External | Cut Ratio | 0.1050 | 0.0588 | 0.1114 | 0.0020 |
| Internal | FOMD | **0.4800** | **1.1651** | **0.3058** | **0.7006** |
| Internal | TPR | 0.2675 | 0.8517 | 0.1665 | 0.5194 |
| Mixed | Conductance | 0.3620 | 0.8128 | 0.0918 | 0.2331 |
| Mixed | Flake ODF | 0.2854 | 0.8384 | 0.1214 | 0.3254 |
| Net-Model | Modularity | 0.0699 | 0.0275 | 0.0675 | 0.0501 |

good performance in small intensities for Expand and large for Shrink, but degrades with larger expansions and smaller reductions. This is further evidence of its resolution limit [FB07].

In this experiment, TPR and FOMD from the internal connectivity family proved to be robust community scoring functions for Twitter interaction streams in a static setting, while Modularity and Cut Ratio proved weaker in the same context. Alternatively, Flake ODF and Conductance – in a lesser degree – from the mixed connectivity family are also reasonably robust choices for microblogging data in the static scenario.

### 4.4.4 Detection Sensitivity

To further complement the investigation of community detection robustness, the sensitivity of the scoring functions in terms of small and large perturbations is explored. For this experiment, the change of Z-score between a small ($p = 0.04$) and a large ($p = 0.20$) perturbation is measured, giving preference to scoring functions that quickly degrade in presence of strong perturbations. The difference $Z(f, h, 0.20) - Z(f, h, 0.04)$ is averaged across all the ground-truth functional communities and the results can be seen in Table 4.3. In these results, large differences indicate that the community scoring function is both robust and sensitive.

In general, the FOMD score (internal connectivity family) stands as the most robust and sensitive score in this experiment for all the perturbation strategies under evaluation in the static scenario. Conductance (mixed connectivity) is the second best for Node Swap, while TPR (internal connectivity) is in all the others. Conversely, the Modularity score performs the worst under every perturbation strategy except Shrink, where only Cut Ratio is worse for microblogging data.

## 4.5 COMMUNITY DETECTION BIAS

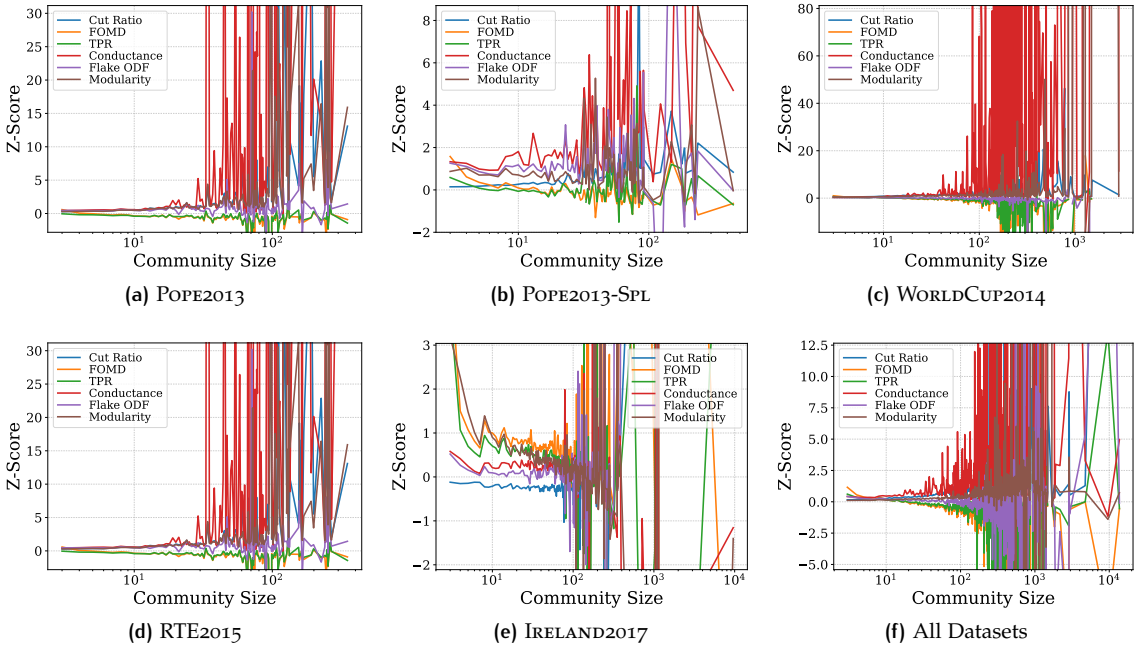
The robustness experiment for the Random perturbation strategy revealed rather large differences in the robustness of the scoring functions. In Figure 4.7, the reported Z-scores for the combined datasets go up to 8.0 standard deviations from the mean, and in the case of the Pope2013-Spl dataset, as high as 15.0 standard deviations from the mean. This suggests that the scoring functions might be subject to a community size bias, where small communities disproportionately affect the results. Therefore, an additional experiment is now proposed to investigate this potential size bias in the scores for the microblogging static scenario.

The experiment is setup as follows. First, a relatively high ($p = 0.20$) constant perturbation intensity is chosen. Then, the changes in the Z-score as a function of the ground-truth community sizes for the selected perturbation intensity is observed. Each Z-score is calculated with respect to all the ground-truth communities with a given size. Because $p = 0.20$ represents a moderately strong intensity for all the investigated perturbation strategies, high values of Z-score that are independent of the community size, i.e. constant, are desired if the scores are in fact unbiased.

Figures 4.10, 4.11, 4.12 and 4.13 respectively show the results for the Node Swap, Random, Shrink and Expand perturbation strategies under the proposed $p = 0.20$ intensity for each ground-truth dataset, including a plot with all the data combined. Initially, the results contained very large Z-score values that subsumed the majority of the smaller values. Therefore, a simple outliers detection strategy [IH93] is applied to the plots with the purpose of improving the visualising of the smaller, more relevant portions of the data.

### 4.5.1 Node Swap

For the **Node Swap** perturbation in Figure 4.10, no scoring function in the experiment is robust for small communities, e.g. with sizes up to $\approx 10^{1.5}$ for the Pope2013, Pope2013-Spl and RTE2015 datasets, and sizes up to $\approx 10^{2}$ for the WorldCup2014 and Ireland2017 datasets. After these size limits, the values of Z-score are much higher. FOMD, TPR and Modularity are the exception to the above observation for the case of Ireland2017, where they exhibit good robustness with smaller communities. This is explained by the long timespan of the dataset, where the ground-truth communities, despite being small, have the most prominent Clustering Coefficient and Cohesiveness structural properties (refer to Table 4.1) among all the datasets.




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

**Figure 4.10:** Z-scores as a function of community size for the Node Swap perturbation strategy applied to all community types for each Twitter ground-truth dataset. A combined plot is also presented.

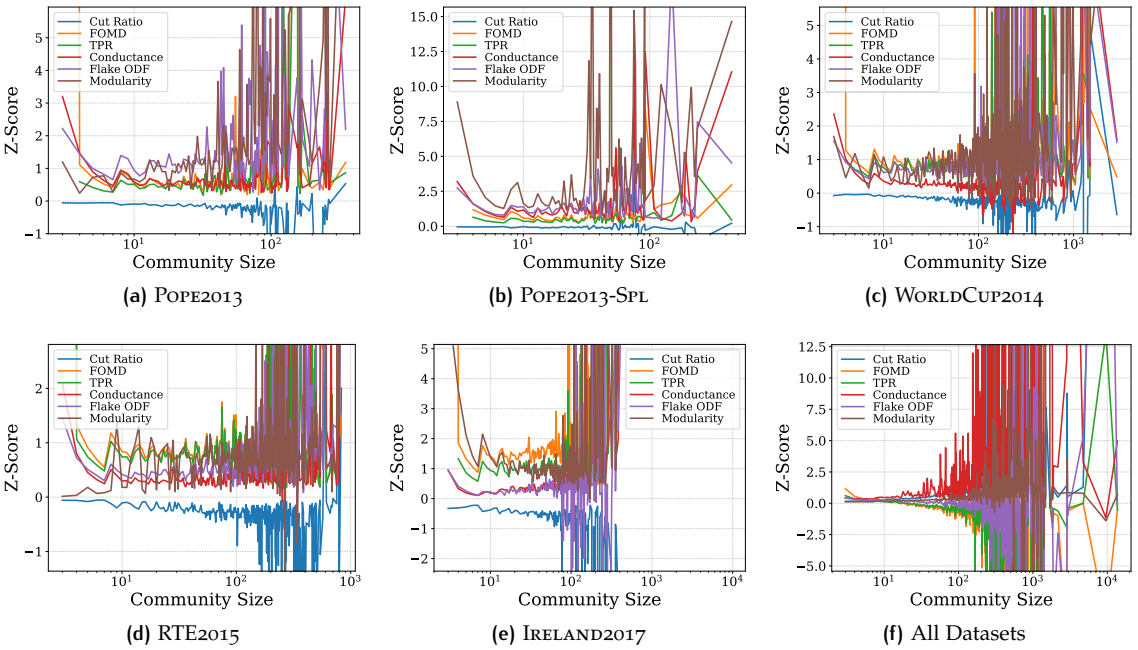
### 4.5.2 Random

For the **Random** perturbation in Figure 4.11, a very similar behaviour is observed. However with this strategy, FOMD, TPR, Conductance and Flake ODF have more consistent robustness across community sizes. Cut Ratio remains stable but with Z-score values close to zero, suggesting that it is not able to distinguish perturbed and non-perturbed communities when the sizes are small enough, e.g. less than $\approx 10^2$ for the POPE2013 and POPE2013-SPL datasets.

### 4.5.3 Expand and Shrink

Lastly, for the **Expand** and **Shrink** perturbations seen in Figure 4.12 and Figure 4.13 also reveal that the scoring functions have a bias for smaller communities, specially in the Expand strategy. In that case, Conductance and Flake ODF (mixed connectivity family) are the more robust in bigger ground-truth communities for the static scenario. On the other hand, for the Shrink perturbation, the Modularity scoring function is prominently more robust on larger communities, again evidencing that its resolution limit also applies to microblogging data.

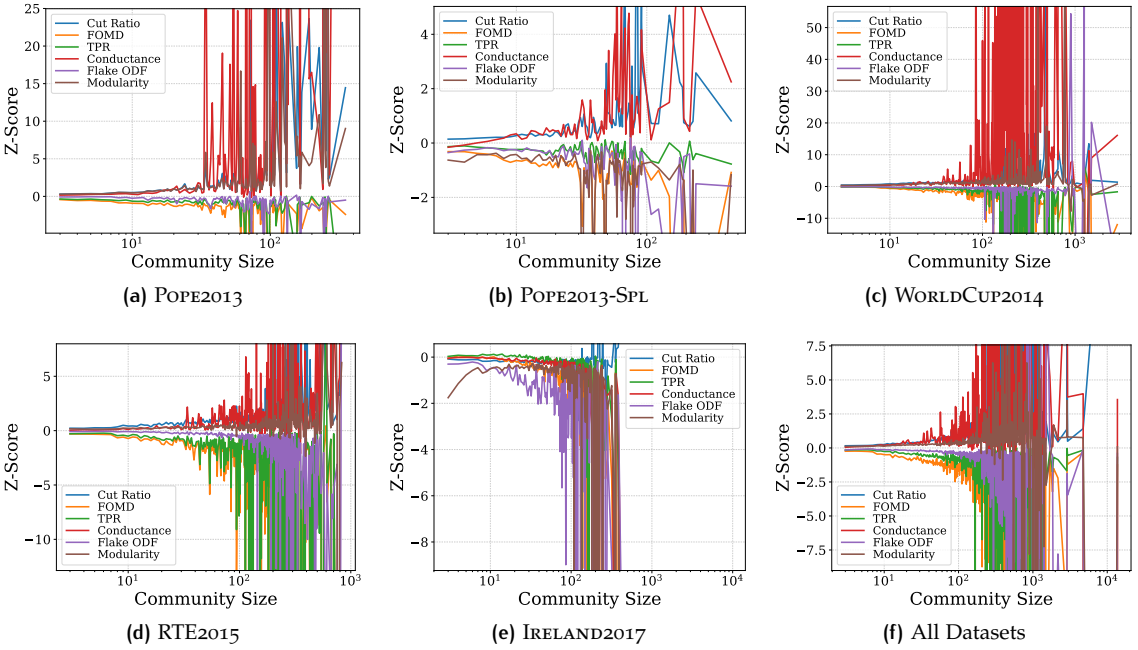
In general, this experiment evidences that all the studied scoring functions have an inherent bias towards small communities, i.e. produce artificially higher performance, for the static




(a) Pope2013 (b) Pope2013-Spl (c) WorldCup2014

(d) RTE2015 (e) Ireland2017 (f) All Datasets

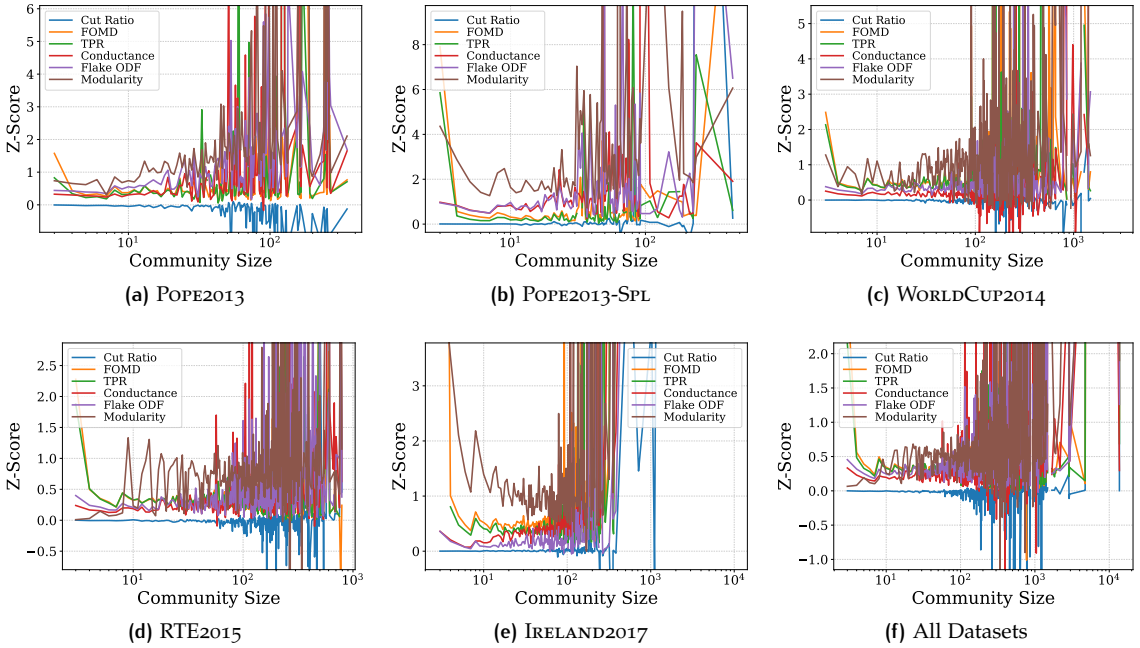
**Figure 4.11:** Z-scores as a function of community size for the Random perturbation strategy applied to all community types for each Twitter ground-truth dataset. A combined plot is also presented.




(a) Pope2013 (b) Pope2013-Spl (c) WorldCup2014

(d) RTE2015 (e) Ireland2017 (f) All Datasets

**Figure 4.12:** Z-scores as a function of community size for the Expand perturbation strategy applied to all community types for each Twitter ground-truth dataset. A combined plot is also presented.




**(a)** POPE2013 **(b)** POPE2013-SPL **(c)** WORLDCUP2014

**(d)** RTE2015 **(e)** IRELAND2017 **(f)** All Datasets

**Figure 4.13:** Z-scores as a function of community size for the Shrink perturbation strategy applied to all community types for each Twitter ground-truth dataset. A combined plot is also presented.

scenario of microblogging social streams. In particular, some of the scores, e.g. Cut Ratio and Modularity, do not perform well when applied to communities smaller than ≈100 users. Nevertheless, the identified size bias does not render the scoring functions incapable of working in microblogging. The scores belonging to the internal and mixed connectivity families proved the most robust and reliable of them all for the analyst to consider, given that the communities under study are large enough. Alternatively, the Conductance and Flake ODF scores are also good candidates for consideration in a lesser degree.

## 4.6 CHAPTER SUMMARY

In this chapter, the problem of evaluating community detection in the context of microblogging services – represented by Twitter – was addressed. First, the structural properties of the constructed functional ground-truth communities in Chapter 3 were evaluated. Afterwards, a set of structural community scoring functions from the literature were thoroughly evaluated using the constructed functional ground-truth in a static scenario that does not consider any temporal information in the microblogging streams. This evaluation investigated the community detection

goodness of the scoring functions and their robustness to a number of perturbation strategies. Furthermore, the sensitivity and bias of the scoring functions were also studied.

For the **static scenario**, the scoring functions based on internal structural information, i.e. from the internal and mixed families, demonstrated to be the best performing. Overall, to identify more clustered, dense and cohesive communities in Twitter, FOMD and TPR are the recommended choices for structural scoring functions. However, if dense but more separated communities are desired by the analyst, then Conductance or Cut Ratio should be considered instead. On the other hand, Modularity and Cut Ratio were found to be weaker in the same context and should not be preferred. As an alternative, Flake ODF and Conductance (in a lesser degree) from the mixed connectivity family were also found reasonably robust for community detection in microblogging data and can be also recommended for consideration.

In terms of robustness to random perturbations, in general the FOMD score (internal connectivity) stands as the most robust and sensitive score for all the perturbation strategies under evaluation in the static scenario. The Modularity score performs the worst under every perturbation strategy except Shrink, where only Cut Ratio is worse for microblogging data.

The experiments in this chapter also evidence that all the studied scoring functions have an inherent bias, i.e. produce artificially higher performance, towards small communities for the static scenario of microblogging social streams. In particular, some of the scores, e.g. Cut Ratio and Modularity, do not perform well when applied to communities smaller than $\approx$100 members. Nevertheless, the identified size bias does not incapacitate the scoring functions in microblogging, but instead this size bias must be taken in consideration.

# 5 TEMPORAL COMMUNITY DETECTION IN MICROBLOGGING

In this chapter, the **dynamic scenario** of community detection in microblogging is investigated. The following main research question, proposed for this scenario in Chapter 1, is addressed.

(RQ3) → *How can activity hotspots based on the dynamic user activity in time be identified in the defined ground-truth communities to improve community detection?*

To provide an answer to this research question, the following research sub-questions are also proposed for this stage, and are investigated in detail in this chapter.

- (RQ3.1) → *What are the temporal characteristics, for instance the user activity distributions, of the defined ground-truth functional communities in (RQ1.1) and (RQ1.2)?*
- (RQ3.2) → *Using the dynamic user activity in time as a basis, how can activity* hotspots *be identified in the defined ground-truth functional communities in (RQ1.1) and (RQ1.2) to be used for further identifying time-scoped sub-communities?*
- (RQ3.3) → *Considering the identified time-scoped sub-communities based on user activity hotspots defined in (RQ3.2), how well do the state-of-the-art structural community definitions investigated in (RQ2.2) now align to these sub-communities in comparison to the ground-truth functional communities in the static scenario, i.e. without considering their user activity context?*

First, a definition for user activity hotspots is introduced and then methods for identifying hotspots in the ground-truth functional communities defined in Chapter 3 are proposed. Afterwards, temporal sub-communities are generated using the identified user activity hotspots and an evaluation is carried in a dynamic scenario of microblogging social networks. The same thirteen structural community definitions discussed in Chapter 4 are re-evaluated using the temporal sub-communities, including their robustness and sensitivity to random perturbations.

The identified contributions of this chapter are: (1) a strategy for the identification of temporal activity hotspots in functional communities in microblogging based on the network of

user interactions, that improves the performance of existing community detection algorithms designed for static data (2) an in-depth characterisation, understanding and evaluation of structural properties for functional communities in microblogging social media, for the dynamic scenario, and (3) a set of recommendations on community detection algorithms based on data-driven evaluation of Twitter user interactions networks.

## 5.1 ACTIVITY HOTSPOTS IN COMMUNITIES

Microblogging social data is fast-pacing and sparse (Chapter 4), therefore the time dimension becomes a fundamental aspect for its analysis and a dynamic scenario where communities are modelled considering their temporal properties must be also investigated.

The construction approach for ground-truth functional communities proposed in Chapter 3 builds each individual community since the first interaction of its members until the last, without any further consideration of the inner activity. However, it is observed in this chapter that the distribution of the user activity in the ground-truth communities is not necessarily uniform. Certain groups of members of a community may become more active than others in the same community at different points in time. Therefore, an extension to the previous ground-truth functional community definition is proposed to investigate the dynamics in time of the communities. In particular, the extended model considers the more active parts of each ground-truth community as individual sub-communities to be analysed independently.

Given a long enough period of time, the network structure of the user interactions associated to functional communities tends to become too dense and, in turn, discovering user communities in these networks becomes more difficult. Therefore, in this thesis it is proposed that, by identifying particular user activity *hotspots* in these networks of interactions, it is possible to find time periods during which user communities are easier to discover.

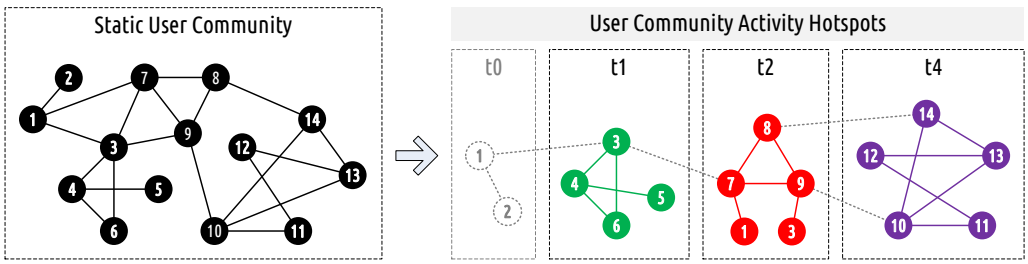



**Figure 5.1:** Example static user community (left) and how it can be divided into temporal sub-communities considering dynamic user activity hotspots in the super-community (right).

An example can be found in Figure 5.1. On the left of the figure, a network of user interactions for an example user community of 14 members can be seen that does not consider any temporal information, i.e. a static scenario. This is the classical representation that many community detection approaches adopt as discussed in Chapter 2. However, in real-world social networks, and specially in microblogging, the user activity between nodes is not equally distributed. Moreover, if network data is captured for a long enough period of time, eventually some links can become much more active than others.

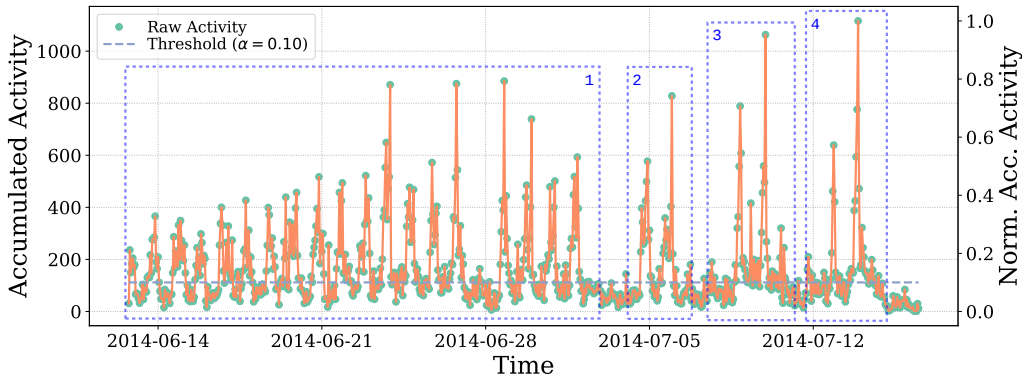
On the other hand, if temporal information is to be considered, i.e. a dynamic scenario, the user activity in time can be used to identify portions of the original network that become more prominent than others. In the right side of the figure, the same network of interactions from the static user community is now divided into example temporal windows according to the user activity recorded in the links, forming temporal sub-communities, i.e. *activity hotspots*, within the static super-community. Note that these temporal windows do not need to be consecutive, i.e. there is no $t_3$ window in the example. Due to low user activity, some links can be discarded (dashed in the figure) and some users can participate in different structures in time, e.g. nodes 1 and 3 in the figure. Moreover, some parts of the static super-community can completely dissolve due to the insufficient size of a sub-community, e.g. nodes 1 and 2 at time $t_0$ in the figure.

In Chapter 4, structural community scoring functions $f(C)$ were introduced, e.g. average degree, fraction over median degree, or modularity. If a user community has varying levels of user activity in time, its structural scores can degrade due to groups of community members also becoming structurally distant in time. However, with the identification of user activity hotspots, this effect can be mitigated because these temporal sub-communities are treated independently and therefore their individual structural scores will not degrade in the same manner.

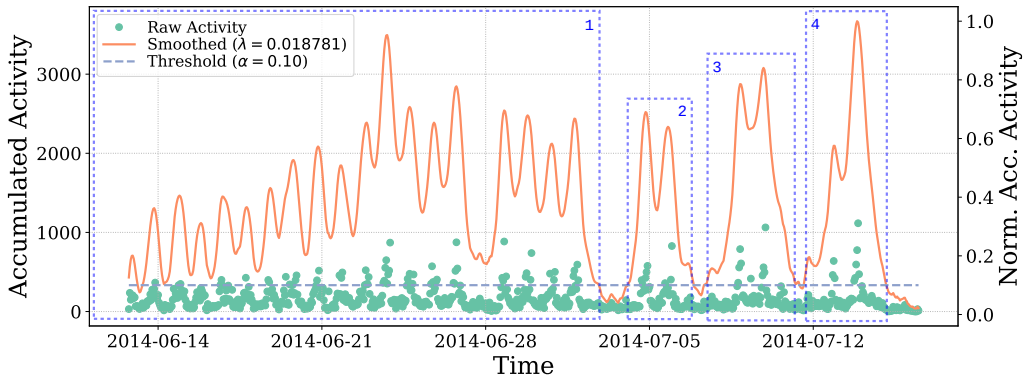
As introduced in Chapter 3, the proposed user interactions network $G = (V, E, W)$ records the temporal user activity using the weighting function $w(e, Q(t, q), type) \in W$, for edges $e = (u_i, u_j) \in E$ between users $u_i, u_j \in V$, where a quantisation parameter $q$ is used to discretise the observed times $t$, e.g. by the minute, hour or day. Using this notation, in this chapter user activity hotspots are defined as follows.

**USER ACTIVITY HOTSPOTS** in a community $C$ are defined as the set $H(C) = \{ts_1, ts_2, \ldots, ts_n\}$ of time spans $ts_i = (t_{start}, t_{end})$ in $C$ where the aggregated user interaction activity at $t_{start}$ is above a defined activation threshold $\alpha$ and at $t_{end}$ is below it. The user interaction activity for $H(C)$ is obtained by accumulating all the weights $w(e, t, type), \forall e = (u_i, u_j) \in E \wedge u_i, u_j \in V : u_i \wedge u_j \in C$, for all interactions occurring at each $t_i$.

First, the aggregated user activity stored in the user interactions network needs to be extracted for each ground-truth functional community C and then user activity hotspots can be identified using a predefined threshold alpha. However, in preliminary experiments it was observed that the raw user activity is not suitable for temporal hotspots analysis due to noise, as seen in Figure 5.2a for an example ground-truth community in the WorldCup2014 dataset.




**(a)** Raw User Activity




**(b)** Smoothed User Activity

**Figure 5.2:** Aggregated user activity per hour (top: raw data, bottom: smoothed using exponential decay) for an example ground-truth community in the WorldCup2014 dataset.

In this example, the raw user activity scale is on the left side of the vertical axis and the $[0, 1]$ min-max scale version for the same data is on the right. Four example activity areas (blue dashed boxes) can be identified in the figure. To attempt to identify these four areas as hotspots, an example threshold $\alpha = 0.10$ for the $[0, 1]$ scale is also observed in the plot. However, in Figure 5.2a too many points transition above and below the proposed threshold during the community lifetime because of the noise in the data. This can be observed in most of the ground-truth communities in all the experimental datasets.

To aid with this data noise problem, an exponential decay smoothing function is adopted. The function activates with aggregated user activity and decays in time during inactive periods based on a time characteristic parameter $\lambda$ [PBV07]. The role of this parameter is to allow for

the activity to cool down at a controlled rate during inactive periods of time. The proposed smoothing function is in Equation 5.1, where the summation runs over all the $n$ recorded raw activity points $a_i$ observed at times $t_i$ for any user community $C$ and allows to compute a smoothed activity value for any arbitrary time $t$.

$$\mathtt{smoothing}(C, t, \lambda) = \sum_{i=0}^{n} a_i e^{-\lambda|t-t_i|/a_i} \tag{5.1}$$

An important property of the proposed smoothing function is that it allows to smooth aggregated user activity measurements for a *continuous* range of times $t$ with any granularity desired, only using the discrete and sparse raw activity observations $a_i$ at times $t_i$.

The proposed smoothing function requires a time characteristic parameter $\lambda$ for the exponential decay component that depends on the activity of every community under study. However, manually choosing a suitable $\lambda$ time characteristic for each ground-truth functional community $C$ is far from practical. Therefore, an estimator for $\lambda$ based on the average of the absolute values of all the adjacent slopes in the raw user activity observations $(t_i, a_i), i \leqslant n$ for any user community $C$ is also proposed and defined in Equation 5.2. It is noted that for this estimator to be properly defined, it requires at least two activity observations in the community $C$. The estimated value of $\lambda$ can be zero because of constant, non-changing activity measurements. In this case, the interpretation for $\lambda$ is that the activity is constant and therefore it does not decay in time during the lifetime of the functional community.

$$\lambda(C) = \frac{\sum_{i=1}^{n} \left| \frac{a_i - a_{i-1}}{t_i - t_{i-1}} \right|}{n-1} \tag{5.2}$$

Returning to the example in Figure 5.2, the smoothing function $\mathtt{smoothing}(C, t, \lambda)$ can be now applied with $\lambda$ estimated using the raw observations in Figure 5.2a to obtain the smoothed user activity version shown in Figure 5.2b for the same example threshold $\alpha = 0.10$. It can be observed that when smoothing the user activity, the four activity areas are much more clear than with the raw activity data. In addition, the smoothed continuous exponential decay curve more gracefully is able to transition above and below the selected example activation threshold $\alpha$, therefore improving the construction of less noisy activity hotspots.

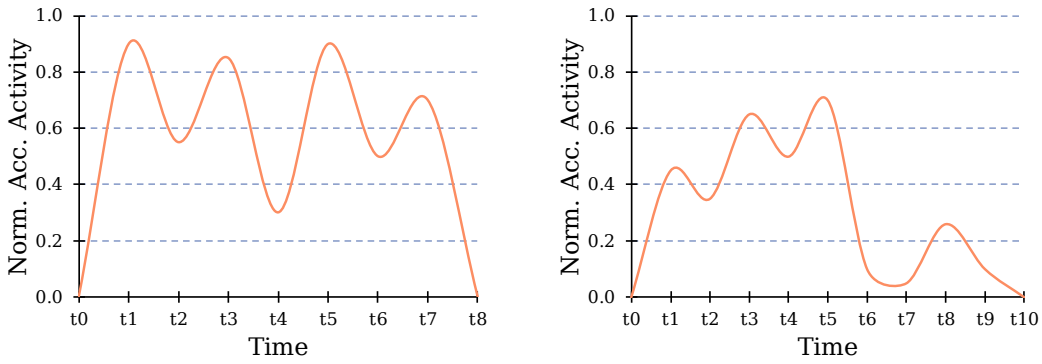
Determining a suitable activation threshold $\alpha$ for the smoothed user activity to be used for detecting activity hotspots and the quality of the resulting temporal sub-communities is the main subject of research in the next sections of this chapter. It is investigated that, given the fast-paced and sparse nature of microblogging user interactions, user communities based on social functions are of better structural quality when focusing on particular portions using their

activity over time. This improvement also translates into better performance for the community detection task using current state of the art community detection approaches.

## 5.2 IDENTIFYING ACTIVITY HOTSPOTS IN COMMUNITIES

After defining user activity hotspots, we next establish an approach for their identification. This task is defined as detecting user activity hotspots $H(C)$ (as defined in Section 5.1) for every ground-truth functional community $C$ in the experimental datasets. For this, first the aggregated user activity is extracted from each ground-truth functional community $C$ and an exponential decay function is applied for smoothing as described in Section 5.1. To obtain a comparable range of values across all the ground-truth communities under study, a min-max normalisation approach in the range $[0, 1]$ to the smoothed activity is also applied.

An activation threshold $\alpha \in [0, 1]$ is now required to find the starting and ending points in time for user activity hotspots $H(C)$ inside each ground-truth community $C$ in the experimental datasets. Each time the normalised aggregated user activity of $C$ rises above $\alpha$, the observed time is recorded as the starting time $t_{start}$ of a new hotspot $ts_i$ in $H(C)$. When the activity falls below $\alpha$, the observed time is recorded as the ending time $t_{end}$ for the same hotspot $ts_i$. Two example user activity patterns – periodic and decaying activities – commonly found in the communities of the experimental datasets can be seen in Figure 5.3.




**(a)** Periodic User Activity

**(b)** Decaying User Activity

**Figure 5.3:** Example patterns of aggregated and normalised user activity smoothed using exponential decay commonly found in ground-truth functional communities. Example evenly-spaced activation thresholds $\alpha$ are shown in dashed blue lines.

In the ideal case of a periodic pattern (Figure 5.3a), as $\alpha$ is raised the average number of hotspots identified should increase and the average number of users inside each hotspot should

reduce because narrower activity windows are being considered. For instance in the example, for $\alpha = 0.2$, 1 hotspot is obtained containing almost all (but not all) the users in the community. For $\alpha = 0.4$, 2 hotspots are now generated with $\approx$45 % of the users in each and $\approx$10 % that is left out, and for $\alpha = 0.6$, 4 very narrow hotspots are created with less users than for $\alpha = 0.4$.

On the other hand, for the case of a decaying pattern (Figure 5.3b), as $\alpha$ is raised the average number of hotspots identified instead remains more constant and the average number of users also constantly reduces. For instance in the example, for $\alpha = 0.2$, 2 hotspots are generated with the majority of the users concentrated on the left-most hotspot. For $\alpha = 0.4$, again 2 hotspots are obtained but with less users within them than before, and for $\alpha = 0.6$, again 2 hotspots can be identified with a very small fraction of users from the original community. Moreover, for $\alpha \approx 0.5$, it is possible to obtain only 1 hotspot concentrating a handful of users.

Choosing the threshold $\alpha$ will depend on the type of communities that the analyst is interested on discovering. However, a systematic method is proposed in this thesis for finding a reasonable threshold for each of the experimental ground-truth datasets to further investigate the performance of community scoring functions in the rest of this chapter. The method is as follows. First, two quantitative metrics are defined for activity hotspots that measure the basic characteristics of a set of generated hotspots given a particular threshold $\alpha$.

#### AVERAGE HOTSPOTS PER COMMUNITY (HPC)

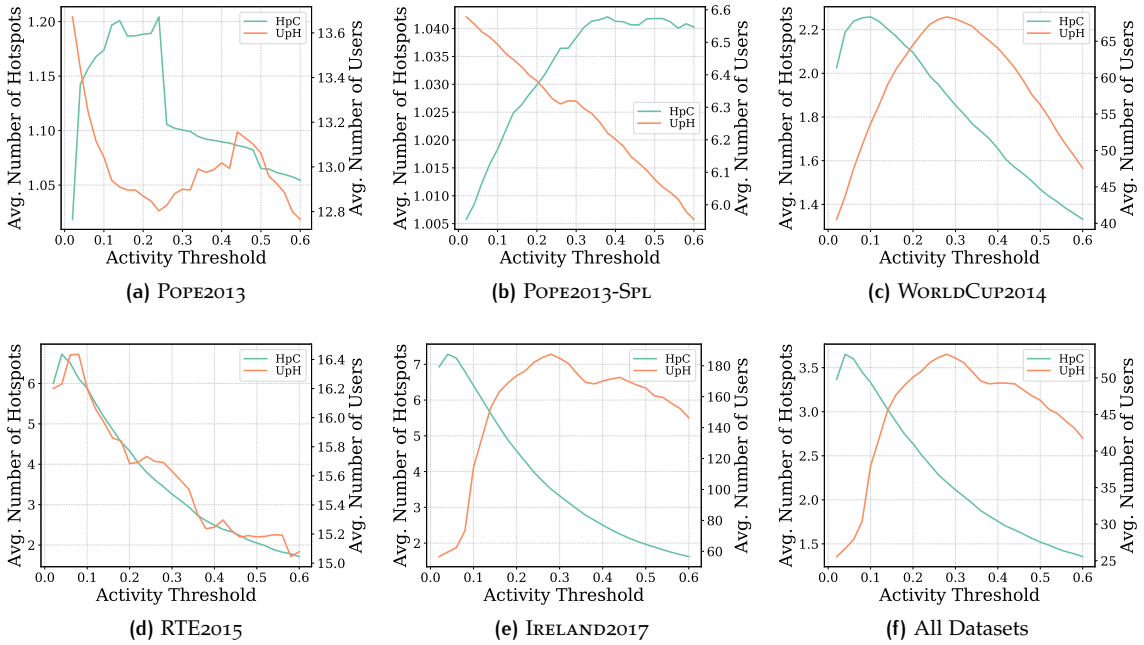
Is the measure of how many user activity hotspots in $H(C)$ are generated per ground-truth functional community $C$ in average. The higher the value of HpC, the more fragmented the original ground-truth communities are becoming.

#### AVERAGE USERS PER HOTSPOT (UPH)

Is the measure of how many users are assigned to each generated user activity hotspot in $H(C)$ in average. The higher the value of UpH, the more packed (or concentrated) the generated hotspots are becoming.

Ideally, a good activation threshold $\alpha$ should jointly maximise HpC and UpH. Therefore the above two metrics will be used as criteria for selecting $\alpha$ within a range of candidates. The experiment is setup by varying $\alpha$ in the range $[0.02, 0.60]$ using increments of 0.02, and measuring HpC and UpC on each set of generated user activity hotspots $H(C)$. This is performed for every community type in all of the experimental datasets. The results can be seen in Figure 5.4, where the range of HpC values is on the left-side and for UpH on the right side of the vertical axis.

In the figures, the maximum average number of generated hotspots per community (HpC) and the maximum average number of users per generated hotspot (UpH) do not coincide in




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

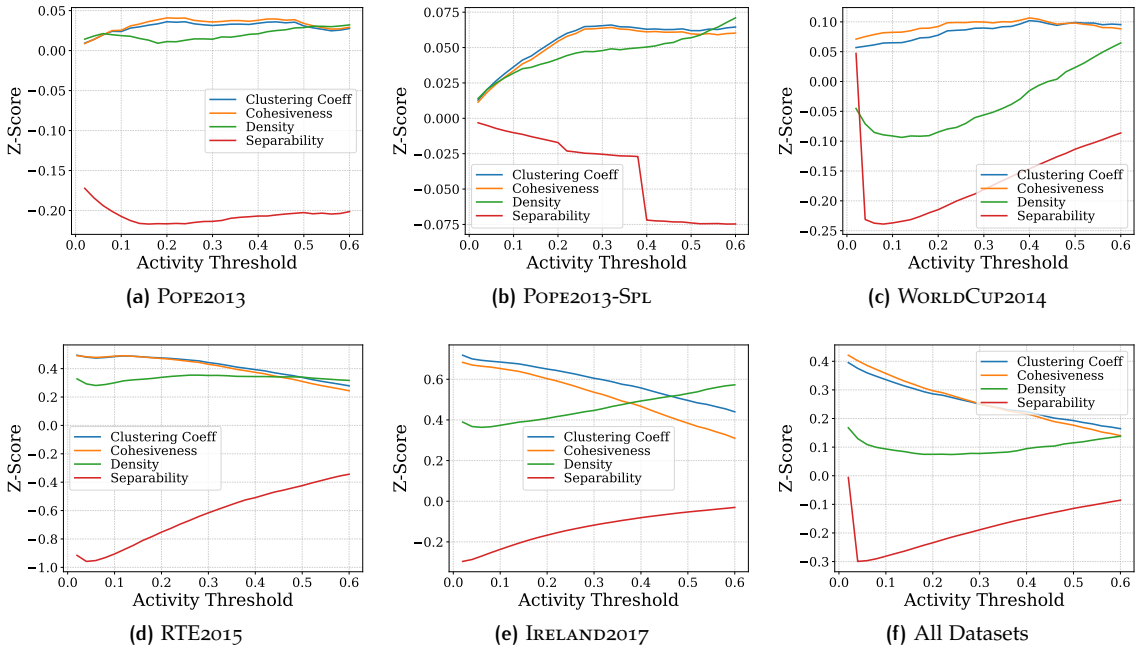
**Figure 5.4:** Activation thresholds $\alpha$ and their effect over the metrics HpC and UpH of the generated activity hotspots $H(C)$ for each Twitter ground-truth dataset. A combined plot is also presented.

any threshold $\alpha$, hence a trade-off must be balanced. The closest maximal coincidence can be observed in the RTE2015 dataset with $\alpha \approx 0.05$.

### 5.2.1 Goodness Metrics of Activity Hotspots

To further complement the previous results, the effect of the activation threshold $\alpha$ over the goodness metrics proposed in Chapter 4 is also investigated. In this proposed experiment, the Z-scores of the difference in goodness metrics between the generated activity hotspots $H(C)$ and the goodness metrics of the source ground-truth communities $C$ for each threshold $\alpha$ in the same range $[0.02, 0.60]$ are observed. Ideally, a good threshold $\alpha$ should maximise all the Z-scores of the goodness metrics under study. The higher the Z-score value, the better is the improvement in goodness when using user activity hotspots. The results of this experiment for each of the ground-truth datasets can be found in Figure 5.5.

It can be observed in the results that the Clustering Coefficient and Cohesiveness goodness metrics always improve when using activity hotspots $H(C)$ in comparison to the complete communities $C$, i.e. their Z-scores is always greater than zero indicating a positive difference in goodness with respect to the mean. This provides preliminary evidence that using activity hotspots enables communities to be more clustered and cohesive and therefore improve their chances




**(a)** POPE2013 **(b)** POPE2013-SPL **(c)** WORLDCUP2014

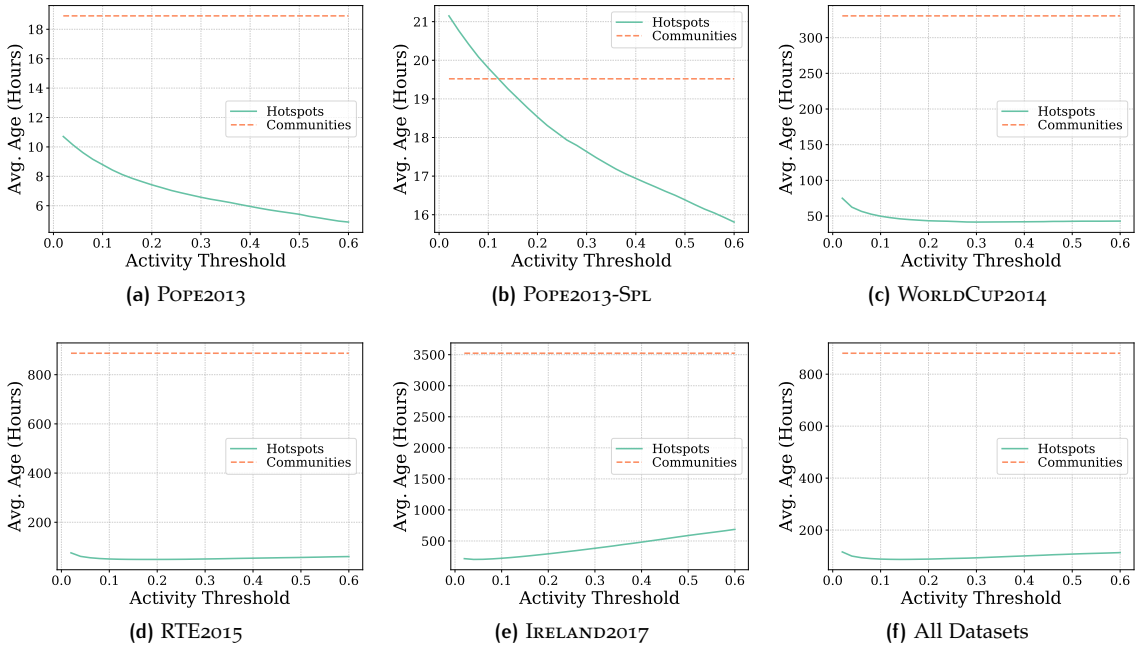
**(d)** RTE2015 **(e)** IRELAND2017 **(f)** All Datasets

**Figure 5.5:** Activation thresholds $\alpha$ and their effect over the four goodness metrics for the generated activity hotspots $H(C)$ for each Twitter ground-truth dataset. A combined plot is also presented.

of being discovered by structural approaches. For the case of the RTE2015 and IRELAND2017 datasets, this improvement is more prominent when the activation threshold $\alpha$ is smaller, i.e. closer to 0.02. On the other hand, for the Papal Event and the WORLDCUP2014 datasets, the improvement in goodness instead is better with a larger value of $\alpha \approx 0.30$, suggesting that in these highly dynamic datasets the activity hotspots are more difficult to separate from inactive periods during the lifetime of the ground-truth communities.

The Density goodness metric displays an opposite behaviour. While Density does exhibit improvement in almost every case (except for the WORLDCUP2014 dataset), its performance does not follow the same trend than the first two goodness metrics studied before. In particular, Density tends to follow an inverse curve as observed in the IRELAND2017 dataset, where the number of generated activity hotspots $H(C)$ decreases but they also become denser. This can be explained by the presence of activity hotspots capturing very narrow activity peaks that concentrate the majority of the users of the original ground-truth community.

Separability is the only metric that never improves when generating activity hotspots $H(C)$ from the ground-truth communities $C$. Moreover, in almost every situation its setback is larger than the improvement presented by the other goodness metrics. This observation is not surprising because every generated user activity hotspot $H(C)$ always contain a subset of the same users in the original community $C$, therefore their separability has high chances to worsen.




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

**Figure 5.6:** Activation thresholds $\alpha$ and their effect over the ages of the generated activity hotspots $H(C)$ for each Twitter ground-truth dataset. The average age of the ground-truth communities $C$ is provided as reference. A combined plot is also presented.

### 5.2.2 Temporal Properties of Activity Hotspots

The ages of the generated user activity hotspots $H(C)$ are also studied. In this experiment, the age of an activity hotspot $H(C)$ is measured as the difference between the start $t_{start}$ and end $t_{end}$ times of the hotspot timespan $ts$ for every activation threshold $\alpha$ in $[0.02, 0.60]$. The results are in Figure 5.6 for each experimental ground-truth dataset, where the average age of the original ground-truth communities $C$ is also provided as reference.

As expected, in average the generated activity hotspots $H(C)$ have much less age than the communities $C$ they originate from. The sole exception is in the POPE2013-SPL dataset, where at $\alpha < 0.15$ the hotspots live longer than the average ground-truth communities. This is due to this particular dataset being random sampled and therefore many ground-truth communities are very short lived. In this case, the identification of user activity hotspots can potentially lead to more meaningful groups of users, i.e. that are engaged for longer time. The hotspots ages also decrease when the activation threshold $\alpha$ is increased because, at higher values of $\alpha$, the hotspots contain less users and therefore less activity. An exception to this observation is the IRELAND2017 dataset, where the hotspots age increase along the threshold $\alpha$ instead of decreasing. This behaviour can be explained by the ground-truth communities having periodic

**Table 5.1:** Selected activity hotspots threshold $\alpha$ for each experimental ground-truth dataset. For each selected threshold, the selection criterion – *average hotspots per community* (HpC) or *average users per hotspot* (UpH) – and its value at the selected threshold is shown. Furthermore, the Z-score of each goodness metric at the same selected threshold is also reported.

| Dataset | $\alpha$ | Criterion | Value | CC | Cohesiv | Density | Separab |
|---|---|---|---|---|---|---|---|
| Pope2003 | 0.24 | HpC | 1.2043 | 0.0360 | 0.0407 | 0.0129 | −0.2165 |
| Pope2003-Spl | 0.28 | UpH | 6.3198 | 0.0650 | 0.0633 | 0.0471 | −0.0250 |
| WorldCup2014 | 0.10 | HpC | 2.2583 | 0.0650 | 0.0825 | −0.0913 | −0.2370 |
| RTE2015 | 0.04 | HpC | 6.7229 | 0.4810 | 0.4836 | 0.2924 | −0.9583 |
| Ireland2017 | 0.04 | HpC | 7.2813 | 0.7003 | 0.6701 | 0.3675 | −0.2866 |

and oscillating changes in user activity, hence when increasing the threshold $\alpha$ the generated hotspots start to merge into longer standing sub-communities.

### 5.2.3 Selecting an Activation Threshold $\alpha$

A simple approach to automatically select an activity threshold $\alpha$ is now proposed based on the average hotspots per community (HpC) and average users per hotspot (UpH) metrics introduced earlier. The proposed method is as follows. For each experimental dataset, the metric that exhibits the higher statistical coefficient of variation – defined as the ratio of the standard deviation $\sigma$ to the mean $\mu$ of the metric values – will be selected as criterion.

Once a criterion metric is selected, a simple signal processing peaks finding algorithm [JOP+01] is applied to identify the global maxima. A summary with all the activity thresholds $\alpha$ selected using this method for all of the community types contained in each of the experimental ground-truth datasets is in Table 5.1. The selected metric (based on highest variation) and its value measured at the chosen threshold are observed.

All the selected thresholds for the experimental ground-truth datasets were found to be $\alpha < 0.30$, suggesting that no more than 30 % of the relative accumulated user activity is required as trigger for forming reasonable hotspots in the ground-truth functional communities. Furthermore, the dominant criterion was found to be HpC, suggesting that the average number of hotspots per community is preferred over the average amount of users in them for selecting good user activity hotspots. The quality of the activity hotspots generated using the selected thresholds is further investigated in the next sections of this chapter.

The method for finding thresholds $\alpha$ presented in this section assumes full visibility in time of all the interactions network activity for each ground-truth community in the experimental datasets under study. However, this assumption is not practical in live microblogging streams where having the complete history of interactions might not be possible. Nonetheless, this

approach can still be applied using a windowing scheme (refer to Section 6.4) where a sliding data window sufficiently large can be adopted to estimate thresholds $\alpha$ dynamically.

## 5.3 STRUCTURAL PATTERNS OF ACTIVITY HOTSPOTS

The user activity hotspots $H(C)$ identified for ground-truth communities $C$ in Section 5.2 allow to construct temporal sub-communities based on the time boundaries of the hotspots. In this thesis it is suggested that these sub-communities $H(C)$ are of better structural quality than the original ground-truth communities $C$. In this section, the structural properties of the activity hotspots $H(C)$ are studied using the same approach described in Section 4.1. In particular, for every constructed activity hotspot sub-community $H(C)$, a corresponding non-community $\widetilde{H(C)}$ is built using random users having similar shortest paths distribution as $H(C)$. Afterwards, the same four structural properties $p$ – Clustering Coefficient, Average Degree, Density and Cohesiveness – are computed for $H(C)$ and $\widetilde{H(C)}$, and the ratio $r = p(H(C))/p(\widetilde{H(C)})$ is observed. As in the original experiment, if this ratio $r$ is greater than 1.0 then there are distinguishable structural patterns in the hotspots $H(C)$ in comparison to randomly chosen nodes with similar shortest path distribution. The results of this new experiment for the same representative experimental datasets, i.e. RTE2015 and IRELAND2017, are found in Table 5.2, where the differences with respect to the static scenario results in Table 4.1 are also observed.

In these representative examples[1], all the structural properties for all the community types show improvement over the base static scenario case, i.e. the differences in ratio $r$ are all positive increments. Moreover, for the RTE2015 dataset, the location-based community types that previously did not exhibit distinguishable structural patterns in the static scenario, now do when using user activity hotspots. This is also observed for the Pope Event datasets. Similar to IRELAND2017 and RTE2015, the WORLDCUP2014 dataset also improved in every community type.

In general, the results of this experiment evidence that, when using accumulated user activity hotspots $H(C)$ for generating temporal sub-communities, these contain more distinguishable structural patterns than the original ground-truth communities $C$ from the static scenario case.

1 All the structural properties can be found in Section B.2 in the appendices.

**Table 5.2:** Ratio between structural properties of user activity hotspots $H(C)$ and randomly chosen nodes with similar shortest path distribution for two representative experimental datasets. The improvement, i.e. differences, with respect to Table 4.1 for each value is also reported.

**(a)** RTE2015

| C. Type | CC | AvgDeg | Density | Cohesiv | All $> 1.0$ |
|---|---|---|---|---|---|
| cities | 5.2716 (+4.87) | 1.2251 (+0.21) | 1.1334 (+0.22) | 4.9728 (+4.42) | Yes (+) |
| countries | 2.4316 (+2.00) | 1.2016 (+0.21) | 1.1400 (+0.19) | 2.4288 (+1.74) | Yes (+) |
| hashtags | 3.5531 (+1.44) | 1.8086 (+0.55) | 1.3510 (+0.26) | 3.0795 (+1.05) | Yes |
| mentions | 4.2429 (+0.48) | 2.2520 (+0.46) | 1.5515 (+0.20) | 3.4606 (+0.29) | Yes |
| places | 3.0567 (+2.66) | 1.1822 (+0.19) | 1.0989 (+0.17) | 3.2423 (+2.76) | Yes (+) |
| quotes | 3.8770 (+1.55) | 1.9087 (+0.52) | 1.5038 (+0.35) | 3.2801 (+1.00) | Yes |
| retweets | 4.0560 (+1.21) | 2.1774 (+0.58) | 1.5220 (+0.34) | 3.3237 (+0.70) | Yes |
| urls | 3.3876 (+0.71) | 1.5076 (+0.21) | 1.2977 (+0.15) | 3.0178 (+0.53) | Yes |
| **Average** | **3.7345** (+1.86) | **1.6579** (+0.37) | **1.3248** (+0.23) | **3.3507** (+1.56) | **Yes** |

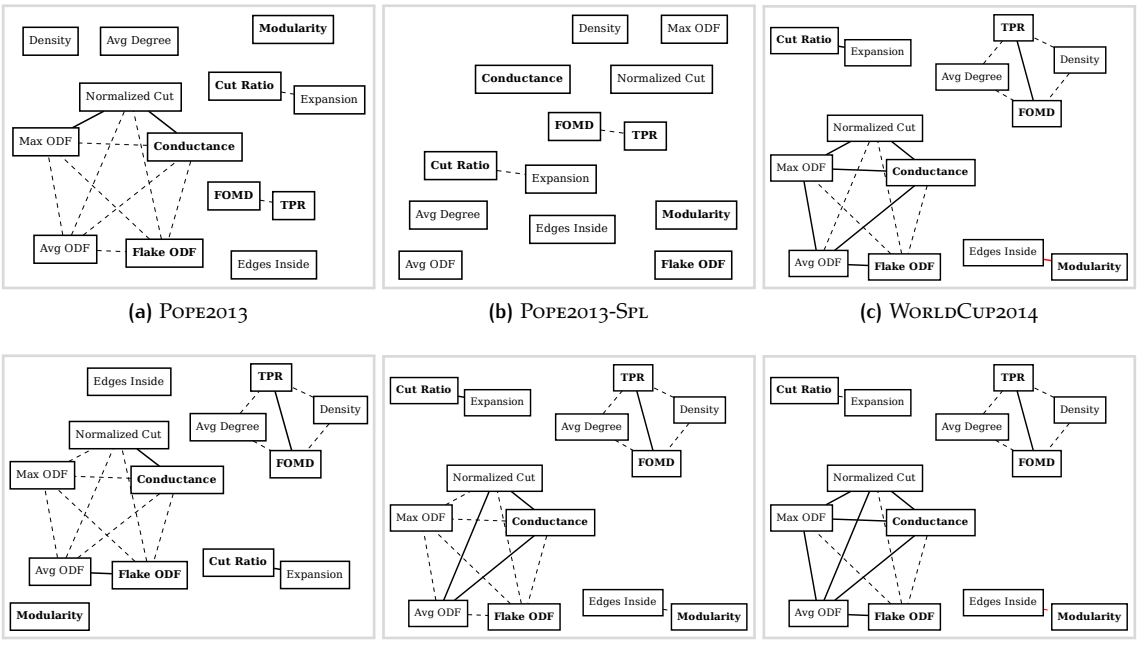
**(b)** IRELAND2017

| C. Type | CC | AvgDeg | Density | Cohesiv | All $> 1.0$ |
|---|---|---|---|---|---|
| cities | 79.8334 (+57.58) | 2.1928 (+1.01) | 1.2812 (+0.21) | 18.3795 (+5.92) | Yes |
| countries | 32.1232 (+23.44) | 2.2361 (+1.21) | 1.0373 (+0.02) | 7.8326 (+0.96) | Yes |
| hashtags | 63.4964 (+31.62) | 1.6238 (+0.40) | 1.3834 (+0.27) | 30.7773 (+15.00) | Yes |
| mentions | 78.9327 (+30.19) | 1.9163 (+0.44) | 1.5747 (+0.34) | 37.6096 (+14.89) | Yes |
| places | 79.6753 (+68.50) | 2.0145 (+0.94) | 1.1689 (+0.13) | 18.2734 (+10.73) | Yes |
| quotes | 68.8420 (+30.59) | 1.5843 (+0.28) | 1.4132 (+0.24) | 34.8451 (+14.50) | Yes |
| **Average** | **67.1505** (+40.32) | **1.9280** (+0.71) | **1.3098** (+0.20) | **24.6196** (+10.34) | **Yes** |

## 5.4 COMMUNITY SCORING FUNCTIONS IN ACTIVITY HOTSPOTS

In Section 4.2, the relationship between the proposed structural scoring functions in the experimental Twitter ground-truth data is explored from a static perspective, i.e. without considering any temporal aspect. In this section, the identified user activity hotspots $H(C)$ are used to generate temporal sub-communities from the ground-truth communities $C$ and the relationship between the same structural scoring functions is re-investigated in this new dynamic scenario.

First, each scoring function $f(H(C))$ applied to each sub-community $H(C)$ generated from the ground-truth functional communities $C$ is computed. Then, the same correlation matrix in Section 4.2 based on the Pearson coefficient is constructed and filtered to unveil the correlation at different degrees between the scoring functions. Again, following the guidelines in [Eva96], $\rho \geqslant 0.3$ and $\rho \geqslant 0.6$ are adopted as thresholds for weak and strong correlation respectively. The results for each experimental dataset and all the ground-truth data combined as well can be seen in Figure 5.7. In the graphs, weak correlation between scores is represented using dashed links and strong correlation with solid connections. All the Pearson coefficients computed for the matrices were found as significant with a small p-value $\leqslant 0.05$.




**(a)** Pope2013 **(b)** Pope2013-Spl **(c)** WorldCup2014

**(d)** RTE2015 **(e)** Ireland2017 **(f)** All Datasets

**Figure 5.7:** Scoring functions applied to the user activity hotspots $H(C)$ generated from communities $C$ and clustered by correlation for each Twitter ground-truth dataset. Weak links ($\rho \geqslant 0.3$) are dashed and strong links ($\rho \geqslant 0.6$) are solid. A combined plot is also presented.

In comparison to the correlations found in Section 4.2 for the static scenario, the following differences can be highlighted. For the Pope2013 dataset, the scoring functions pair (Cut Ratio, Expansion) now correlates weakly and the pairs (Max ODF, Normalized Cut) and (Normalized Cut, Conductance) now correlate strongly instead of weakly. For the Pope2013-Spl dataset, the pairs (Cut Ratio, Expansion) and (FOMD, TPR) now correlate weakly, in contrast to a total absence of correlation in the static case. These observations suggest that in the dynamic case using activity hotspots, the scoring functions belonging to the same families can better correlate.

For the larger WorldCup2014, RTE2015 and Ireland2017 datasets, the internal connectivity scoring functions now exhibit a stronger correlation between each other: Avg. Degree, Density, TPR and FOMD. However, the Edges Inside score, also from the internal connectivity family, still remains an outsider. Moreover, the Edges Inside and Modularity scores now correlate weakly in the Ireland2017 dataset but unexpectedly correlate *inversely* in the WorldCup2014 dataset. These observations suggest that the number of edges inside a community is an unreliable internal connectivity metric for the dynamic scenario.

The following scoring functions from the mixed family also become more strongly correlated in the WorldCup2014 dataset: Max ODF, Avg ODF, Normalized Cut and Conductance. However they also become weaker in the Ireland2017 dataset. Furthermore, only Max ODF and Avg ODF

become weaker in the RTE2015 dataset. Lastly, the external connectivity scores Cut Ratio and Expansion also correlate more strongly in the dynamic scenario using activity hotspots.

In general, all of the scoring functions correlate even more into their four predefined families when using user activity hotspots in the dynamic scenario compared to the time-independent static case for Twitter. Therefore, the same six representative scoring functions defined in Section 4.2 and also in bold in Figure 5.7 will remain unchanged in this chapter: FOMD, TPR, Cut Ratio, Conductance, Flake ODF and Modularity.

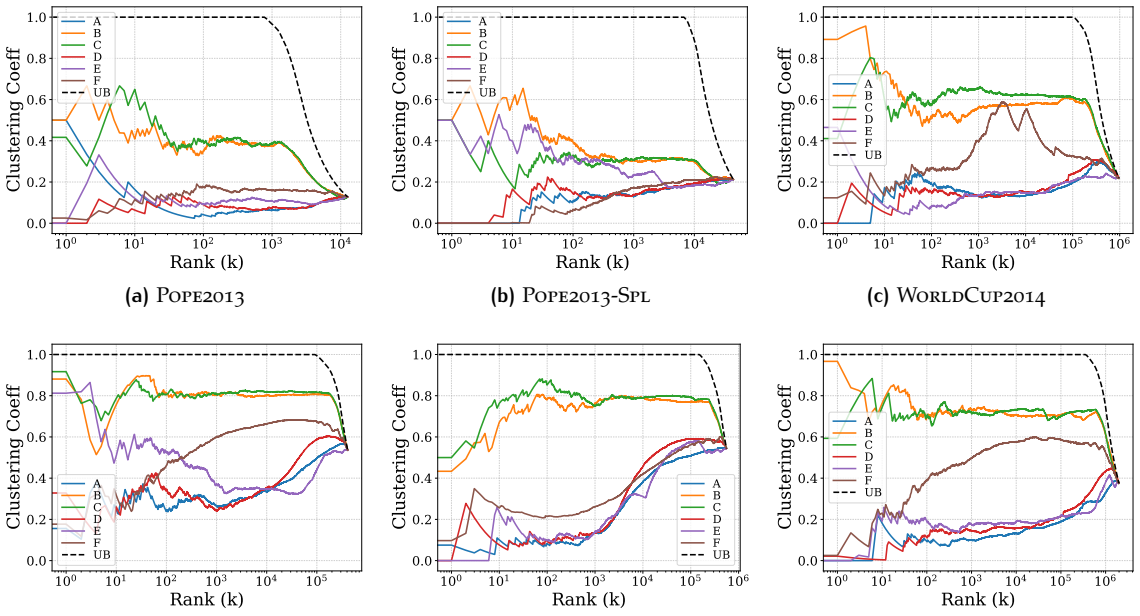
## 5.5 GOODNESS OF ACTIVITY HOTSPOTS DETECTION

The goodness metrics from Section 4.3 applied to the newly formed temporal sub-communities $H(C)$ using activity hotspots is now investigated. The goal is to demonstrate that, when considering user activity for the identification of activity hotspots, the resulting sub-communities $H(C)$ are of better quality when compared to the originating communities $C$. Evidence of this is shown by measuring the goodness of the activity hotspots generated from the experimental ground-truth functional communities and comparing it to the same goodness metrics previously discussed in the static scenario evaluation.

To evaluate the goodness of the scoring functions for the ground-truth user activity hotspots $H(C)$, the same experiment from Section 4.3 is adopted. For each dataset and community type, the temporal sub-communities $H(C)$ are ranked using the six selected scoring functions $f(H(C))$ in descending order. Then, the cumulative moving average (CMA) of each goodness metric $g(H(C))$ is observed for the top-$k$ ground-truth communities under the order induced by $f(H(C))$. A perfect scoring function should rank the temporal sub-communities in the same descending order as the goodness metrics, and therefore the CMA should decrease monotonically along $k$. Conversely, a poor community scoring function would produce a $k$-dependent constant CMA.

Figures 5.8, 5.9, 5.10 and 5.11 respectively show the results for the ranked Clustering Coefficient, Density, Cohesiveness and Separability for each ground-truth dataset, including a plot with all the data combined. The upper bound curve, i.e. the CMA of a goodness metric ranked by the same goodness metric, which represents a perfect ranking is also provided for reference.

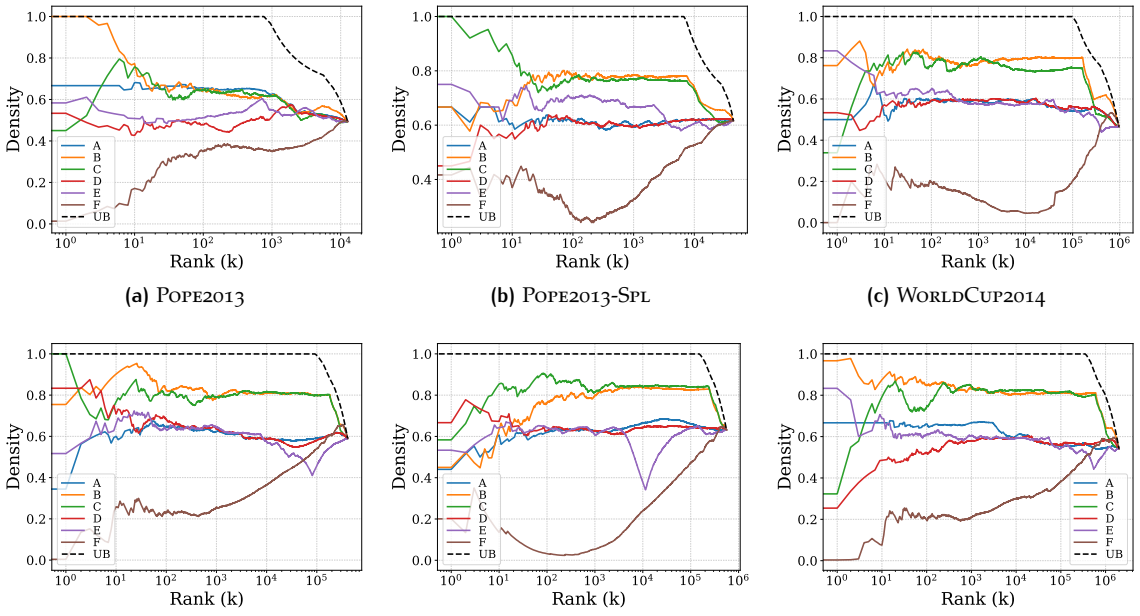
To complement the results in the figures and provide a better insight about the quality of the goodness ranking of the scoring functions for the generated temporal sub-communities $H(C)$, the following experiment is proposed. It was previously stated that a perfect scoring function should rank the temporal sub-communities in the same descending order as the goodness met-




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

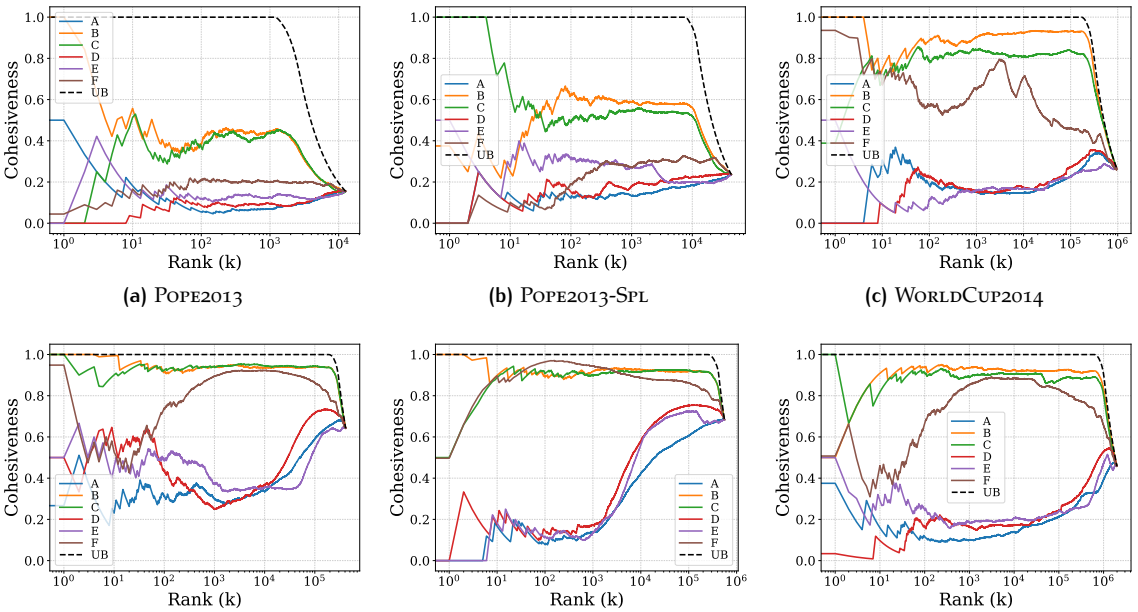
**Figure 5.8:** Ranked Clustering Coefficient by CMA of temporal sub-communities $H(C)$ based on user activity hotspots generated from communities C for each Twitter ground-truth dataset. A combined plot is also presented. Scores: Cut Ratio (A), FOMD (B), TPR (C), Conductance (D), Flake ODF (E), Modularity (F) and their Upper Bound (UB).




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

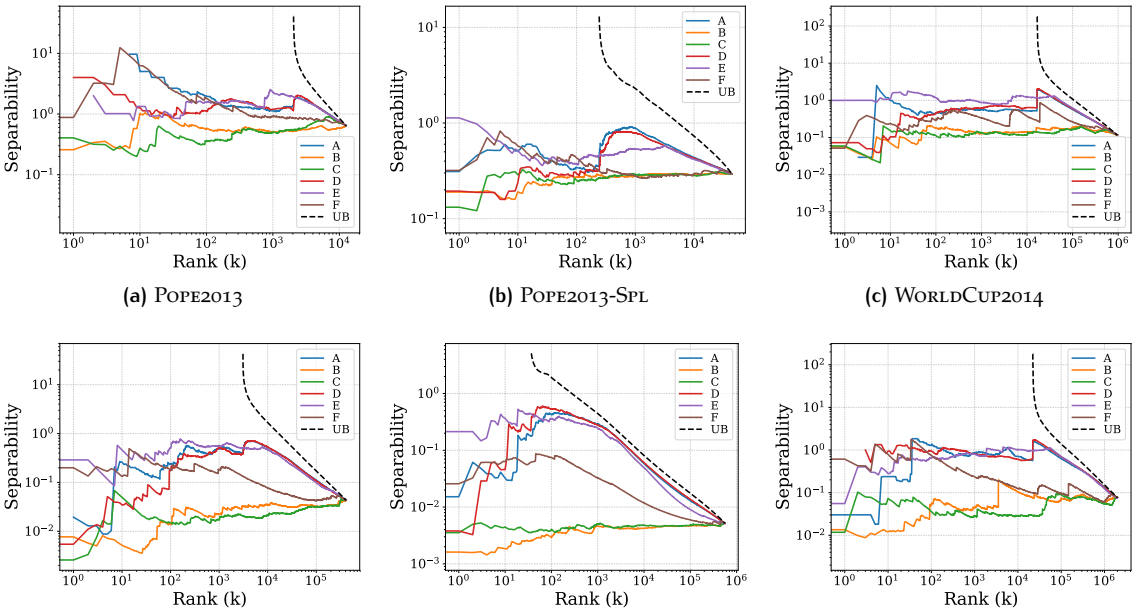
**Figure 5.9:** Ranked Density by CMA of temporal sub-communities $H(C)$ based on user activity hotspots generated from communities C for each Twitter ground-truth dataset. A combined plot is also presented. Scores: Cut Ratio (A), FOMD (B), TPR (C), Conductance (D), Flake ODF (E), Modularity (F) and their Upper Bound (UB).




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

**Figure 5.10:** Ranked Cohesiveness by CMA of temporal sub-communities $H(C)$ based on user activity hotspots generated from communities C for each Twitter ground-truth dataset. A combined plot is also presented. Scores: Cut Ratio (A), FOMD (B), TPR (C), Conductance (D), Flake ODF (E), Modularity (F) and their Upper Bound (UB).




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

**Figure 5.11:** Ranked Separability by CMA of temporal sub-communities $H(C)$ based on user activity hotspots generated from communities C for each Twitter ground-truth dataset. A combined plot is also presented. Scores: Cut Ratio (A), FOMD (B), TPR (C), Conductance (D), Flake ODF (E), Modularity (F) and their Upper Bound (UB).

**Table 5.3:** Average Z-score distances of the ranked scoring functions with respect to the upper bound for the Clustering Coefficient goodness metric in the static and dynamic scenarios. Improvements in the dynamic case are marked with ↑. Values closer to zero are considered better.

| **Static** | Cut Ratio | FOMD | TPR | Conduct | Flake ODF | Modularity |
|---|---|---|---|---|---|---|
| Pope2013 | −1.0152 | −0.6227 | −0.8331 | −0.9617 | −0.7280 | −0.7382 |
| Pope2013-Spl | −0.7862 | −0.7391 | −0.7543 | −0.8091 | −0.7932 | −0.7858 |
| WorldCup2014 | −0.9982 | −0.9517 | −0.9245 | −0.9914 | −0.9885 | −0.9560 |
| RTE2015 | −1.4071 | −1.1451 | −1.0971 | −1.3478 | −1.3513 | −1.1468 |
| Ireland2017 | −1.0778 | −0.9693 | −0.9513 | −1.0774 | −1.0622 | −1.0065 |
| **Dynamic** | Cut Ratio | FOMD | TPR | Conduct | Flake ODF | Modularity |
| Pope2013 | ↑ −0.9414 | −0.8278 | ↑ −0.7573 | ↑ −0.7980 | −0.7807 | ↑ −0.7344 |
| Pope2013-Spl | −0.8403 | −0.7891 | −0.7852 | −0.8391 | −0.8271 | −0.8302 |
| WorldCup2014 | −1.0573 | ↑ −0.7342 | ↑ −0.7319 | −1.0267 | −1.0904 | −0.9589 |
| RTE2015 | −1.9386 | ↑ −0.8699 | ↑ −0.8575 | −1.7080 | −2.1102 | −1.4629 |
| Ireland2017 | −1.6096 | ↑ −0.8281 | ↑ −0.7907 | −1.4178 | −1.5952 | −1.3929 |

rics, therefore all the goodness ranking curves should align to the upper bound (UB) curve. It is then possible to assess the quality of the ranked goodness by measuring how distant are the rankings to their upper bound. This distance will be measured by computing the Z-score of each scoring function with respect to the upper bound curve, i.e. Equation 5.3 for each goodness metric and average this value over all the ranks $k$, i.e. $\mathrm{ubDist}(f) = \sum Z(f)/k$.

$$Z(f) = \frac{E[g(H(C_i)) - f(H(C_i))]}{\sqrt{Var[g(H(C_i))]}} \tag{5.3}$$

If a scoring function $f_i$ is better ranked in terms of a particular goodness metric than another scoring function $f_j$, then $\mathrm{ubDist}(f_i) < \mathrm{ubDist}(f_j)$. Also, because $\mathrm{ubDist}(f)$ measures the distance of a scoring function $f$ to the upper bound of the goodness metric, its value is always negative and being closer to zero, i.e. less negative, is considered better. The results of this experiment for each ground-truth experimental dataset in both, the static (without using activity hotspots) and the dynamic scenarios (using activity hotspots), can be found in Table 5.3 for the Ranked Goodness, Table 5.4 for the Density, Table 5.5 for the Cohesiveness and Table 5.6 for the Separability goodness metrics. The symbol ↑ indicates an improvement in the ranking quality.

In comparison to the static scenario, all of the scoring functions are able to improve their ranking alignment to the goodness metrics at different degrees when considering the proposed activity hotspots $H(C)$ for generating temporal sub-communities from the ground-truth communities $C$. The improvement is evident in the larger datasets, i.e. WorldCup2014, RTE2015 and Ireland2017, where the FOMD (B) and TPR(C) scoring functions now achieve a closer alignment to the upper bounds (UB) than before in every dataset and goodness metric. Moreover, in terms of the Separability goodness metric, every scoring function improves its ranking in these

**Table 5.4:** Average Z-score distances of the ranked scoring functions with respect to the upper bound for the Density goodness metric in the static and dynamic scenarios. Improvements in the dynamic case are marked with ↑. Values closer to zero are considered better.

| **Static** | Cut Ratio | FOMD | TPR | Conduct | Flake ODF | Modularity |
|---|---|---|---|---|---|---|
| Pope2013 | −1.3228 | −1.0544 | −1.2378 | −1.6137 | −1.7619 | −1.8983 |
| Pope2013-Spl | −1.9700 | −2.2509 | −2.1202 | −2.1173 | −2.5768 | −3.2064 |
| WorldCup2014 | −1.4719 | −1.0621 | −1.5649 | −1.4900 | −1.5865 | −2.5584 |
| RTE2015 | −1.6090 | −1.7473 | −1.7943 | −1.6860 | −1.6250 | −2.4174 |
| Ireland2017 | −1.3910 | −1.6022 | −1.5674 | −1.4406 | −1.4993 | −1.9785 |
| **Dynamic** | Cut Ratio | FOMD | TPR | Conduct | Flake ODF | Modularity |
| Pope2013 | ↑ −1.1672 | −1.1488 | −1.6230 | ↑ −1.1221 | ↑ −1.4237 | −2.2233 |
| Pope2013-Spl | −2.0547 | −2.3352 | ↑ −1.7868 | −2.4358 | ↑ −2.1970 | −3.4988 |
| WorldCup2014 | ↑ −1.2835 | ↑ −0.8928 | ↑ −1.2540 | ↑ −1.2386 | ↑ −1.5043 | ↑ −2.2111 |
| RTE2015 | −1.6529 | ↑ −0.9679 | ↑ −0.9852 | ↑ −1.6780 | −2.1020 | ↑ −1.9428 |
| Ireland2017 | ↑ −1.2282 | ↑ −1.2487 | ↑ −1.2113 | ↑ −1.3232 | −1.5193 | −2.4974 |

**Table 5.5:** Average Z-score distances of the ranked scoring functions with respect to the upper bound for the Cohesiveness goodness metric in the static and dynamic scenarios. Improvements in the dynamic case are marked with ↑. Values closer to zero are considered better.

| **Static** | Cut Ratio | FOMD | TPR | Conduct | Flake ODF | Modularity |
|---|---|---|---|---|---|---|
| Pope2013 | −1.2239 | −0.6774 | −0.7925 | −1.3027 | −0.7342 | −0.7046 |
| Pope2013-Spl | −0.8863 | −0.5658 | −0.5788 | −0.8547 | −0.8719 | −0.7433 |
| WorldCup2014 | −1.1292 | −0.4116 | −0.7426 | −1.0913 | −1.0798 | −0.8922 |
| RTE2015 | −1.5954 | −1.1020 | −0.9270 | −1.5173 | −1.7266 | −1.3303 |
| Ireland2017 | −1.4135 | −0.9822 | −0.9753 | −1.3337 | −1.3609 | −1.1614 |
| **Dynamic** | Cut Ratio | FOMD | TPR | Conduct | Flake ODF | Modularity |
| Pope2013 | ↑ −1.0585 | −0.8738 | ↑ −0.6650 | ↑ −1.0136 | −0.8986 | −0.9085 |
| Pope2013-Spl | −0.9755 | ↑ −0.5578 | ↑ −0.5736 | −0.8972 | −0.9196 | ↑ −0.7418 |
| WorldCup2014 | −1.2563 | ↑ −0.2337 | ↑ −0.4062 | −1.1718 | −1.2775 | ↑ −0.8540 |
| RTE2015 | −2.6583 | ↑ −0.5413 | ↑ −0.4912 | −2.1131 | −2.9056 | ↑ −1.3165 |
| Ireland2017 | −2.7523 | ↑ −0.6478 | ↑ −0.6366 | −2.0500 | −2.4012 | ↑ −0.9738 |

**Table 5.6:** Average Z-score distances of the ranked scoring functions with respect to the upper bound for the Separability goodness metric in the static and dynamic scenarios. Improvements in the dynamic case are marked with ↑. Values closer to zero are considered better.

| **Static** | Cut Ratio | FOMD | TPR | Conduct | Flake ODF | Modularity |
|---|---|---|---|---|---|---|
| Pope2013 | −0.4774 | −0.9419 | −0.6515 | −0.5490 | −0.9288 | −0.9609 |
| Pope2013-Spl | −0.5136 | −0.7593 | −0.7578 | −0.4658 | −0.5838 | −0.6285 |
| WorldCup2014 | −0.3402 | −0.4318 | −0.4423 | −0.3178 | −0.3592 | −0.4580 |
| RTE2015 | −0.5604 | −0.6286 | −0.6039 | −0.5578 | −0.5019 | −0.6173 |
| Ireland2017 | −0.3483 | −0.4598 | −0.4619 | −0.3347 | −0.3704 | −0.4516 |
| **Dynamic** | Cut Ratio | FOMD | TPR | Conduct | Flake ODF | Modularity |
| Pope2013 | −0.7252 | ↑ −0.8337 | −0.6579 | −0.6640 | ↑ −0.7445 | ↑ −0.7111 |
| Pope2013-Spl | ↑ −0.5042 | −0.7834 | ↑ −0.6902 | −0.5225 | −0.6158 | −0.7095 |
| WorldCup2014 | ↑ −0.2875 | ↑ −0.3470 | ↑ −0.3527 | ↑ −0.2703 | ↑ −0.2874 | ↑ −0.3767 |
| RTE2015 | ↑ −0.2563 | ↑ −0.3936 | ↑ −0.3870 | ↑ −0.2459 | ↑ −0.3829 | ↑ −0.3587 |
| Ireland2017 | ↑ −0.2039 | ↑ −0.3936 | ↑ −0.3964 | ↑ −0.1891 | ↑ −0.3111 | ↑ −0.3874 |

Table 5.7: Aggregated scoring ranking by goodness metrics using the Borda voting method for all ground-truth datasets. Best ranked scoring functions for each goodness metric are in bold. The best scores in the static scenario (Table 4.2) are marked with ⋄.

| Family | Score | CC | Cohesiveness | Density | Separability |
|---|---|---|---|---|---|
| External | Cut Ratio | 5.4582 | 5.4529 | 4.7199 | **1.4838** |
| Internal | FOMD | **1.6001** | ⋄ **1.0816** | ⋄ **1.1347** | 4.9085 |
| Internal | TPR | ⋄ 1.7027 | 2.0805 | 2.6169 | 5.9127 |
| Mixed | Conductance | 3.5261 | 3.6173 | 3.2281 | ⋄ 1.9035 |
| Mixed | Flake ODF | 5.3835 | 5.4022 | 5.7020 | 3.1322 |
| Net-Model | Modularity | 3.3294 | 3.3654 | 3.5985 | 3.6593 |

large datasets. This is an unexpected result because Separability did not previously show an improvement in Section 5.2 while selecting an activation threshold. This suggests that, despite the Separability of the temporal sub-communities $H(C)$ not improving using hotspots, the scoring functions are still better ranked in terms of Separability. The Conductance score improves in terms of Density and Separability, but not in Cohesiveness and Clustering Coefficient. Another improvement can be seen in Modularity (F) for the Cohesiveness and Separability goodness metrics, as now it is able to attain similar performance than FOMD and TPR.

In general, the internal and mixed connectivity family of scores are the ones benefiting the most from using user activity hotspots to identify temporal sub-communities in the ground-truth. Nevertheless, the other families also can benefit in certain cases. Overall, the activity hotspots are able to improve the performance of structural scoring functions in terms of Cohesiveness, Density and Clustering Coefficient, but not Separability, which remains mostly unaffected.

### 5.5.1 Goodness Metrics Ranking

The aggregated scoring ranking by goodness metric studied in Section 4.3 is also re-investigated for the temporal sub-communities $H(C)$ generated using activity hotspots. Again, for each goodness metric $g(H(CC))$ and scoring function $f(H(C))$ in all of the ground-truth datasets, the rank of each score in comparison to the other scoring functions at every rank $k$ is observed. The six scores are ranked and aggregated using the Borda voting method [Saa12] to obtain an unified ranking that quantifies the ability of each scoring function to find *good communities*. The results for all the ground-truth datasets combined are in Table 5.7, where ranks $\approx 1.0$ (in bold in the table) indicate scoring functions adequate for each goodness criteria and the previous bests from Section 4.3 are marked with ⋄.

Similar to the previous results in the static scenario, the representatives using internal structural information demonstrated to be the dominant performing in this experiment as well, with

the exception of Cut Ratio, which outperformed Conductance in terms of Separability in the dynamic scenario. One minor change observed is that the FOMD scoring function is now slightly preferred over TPR in terms of better Clustering Coefficient.

Overall, the goodness experiments in this section suggest that, to identify more clustered, dense and cohesive communities in Twitter in a time-aware context, FOMD and TPR are the better choices for structural scoring functions. If dense but more separated communities are desired by the analyst, then Conductance or Cut Ratio should be considered.

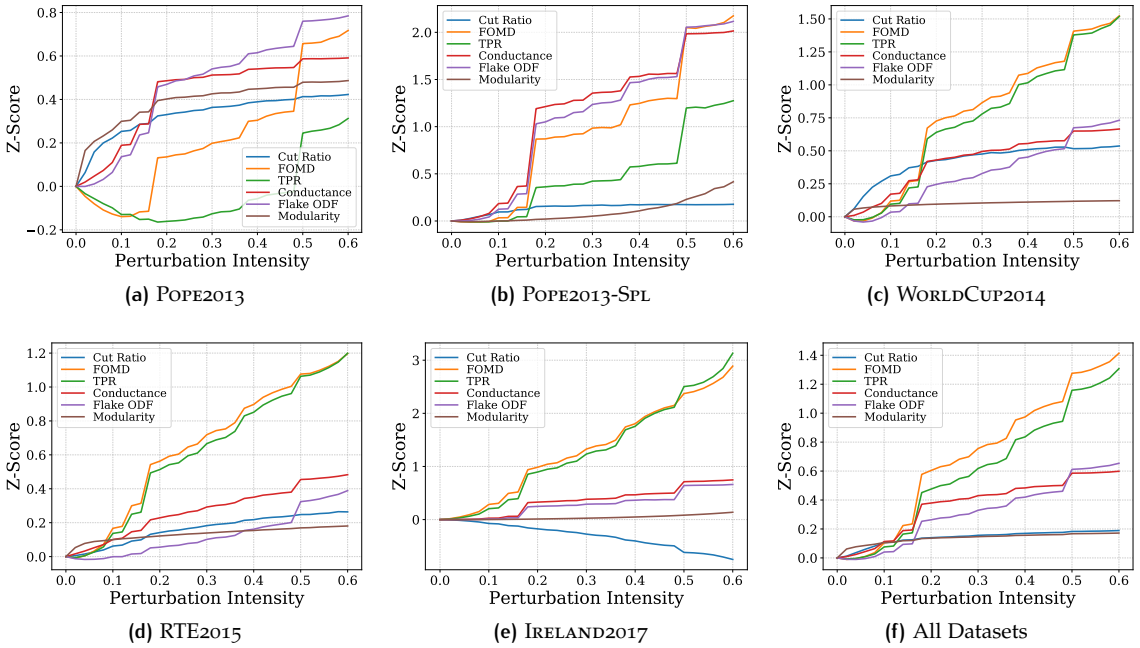
## 5.6 ROBUSTNESS OF ACTIVITY HOTSPOTS DETECTION

The robustness of the scoring functions in the context of the temporal sub-communities generated using user the proposed activity hotspots $H(C)$ is now investigated. Good scoring functions should be stable under small perturbations and reduce their performance under strong disturbance. In Section 4.4, a set of community perturbation strategies were proposed for studying the robustness of the structural scoring functions: Node Swap, Random, Expand and Shrink. In this section, these strategies are now applied to the temporal sub-communities $H(C)$ generated using activity hotspots and compared to the static scenario.

The perturbation experiment for the dynamic scenario is now defined as follows. Similar as in Section 4.4, the perturbation intensity is varied in the range $p \in [0.01, 0.60]$, e.g. in the Node Swap strategy this means exchanging between 1 and 60 % of the members of a community, and observe the averaged Z-score across all ground-truth temporal sub-communities $H(C)$ in all community type and datasets. Figures 5.12, 5.13, 5.14 and 5.15 respectively show the averaged Z-score results for the Node Swap, Random, Shrink and Expand perturbation strategies under the proposed intensities for each ground-truth dataset, including a plot with all the data combined.

### 5.6.1 Node Swap

For the **Node Swap** perturbation in Figure 5.12, similar to the static case the TPR and FOMD scores perform the best in all the long timespan datasets (WorldCup2014, RTE2015 and Ireland2017), followed by Conductance and Flake ODF. In the case of the Pope Event datasets, Conductance and Flake ODF instead are observed as more robust scores. In contrast, Modularity




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

**Figure 5.12:** Z-scores of intensities for the Node Swap perturbation strategy applied to all temporal sub-communities $H(C)$ based on user activity hotspots generated from communities $C$ for all community types in each Twitter ground-truth dataset. A combined plot is also presented.

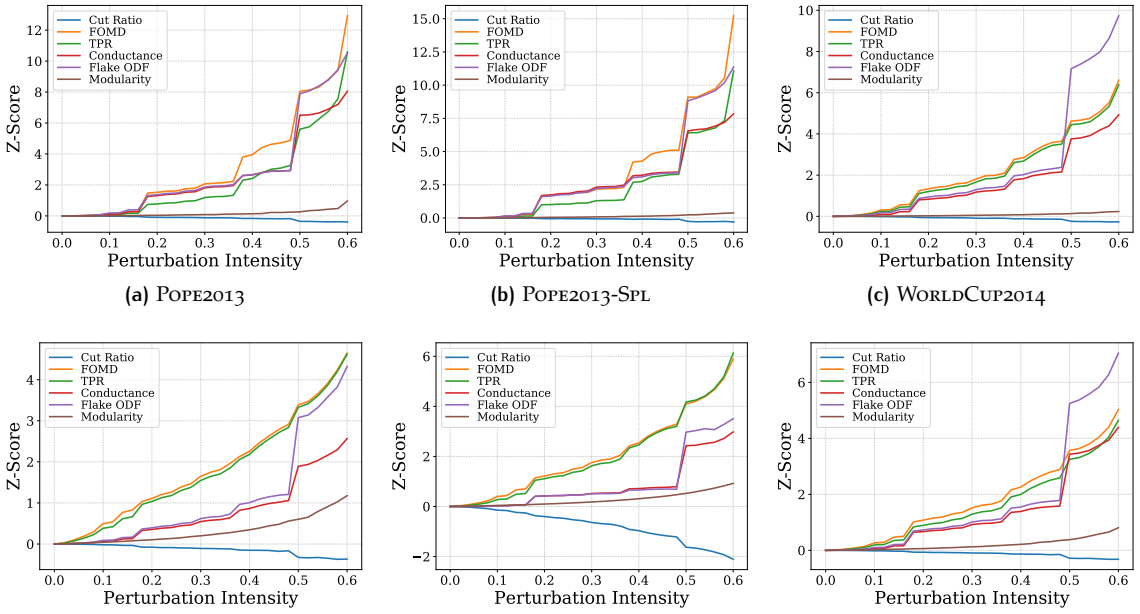
and Cut Ratio do not degrade as gracefully – particularly Modularity – when the perturbation is increased, revealing their inability to handle noisy data in Twitter also for the dynamic scenario.

### 5.6.2 Random

For the **Random** perturbation in Figure 4.7, again the results are very similar as in static case. The internal and mixed connectivity families of scores – FOMD, TPR, Conductance and Flake ODF – consistently perform the best, with the internal family being very robust (for example in the IRELAND2017 dataset). Cut Ratio and Modularity still perform poorly in in presence of strong noise in the dynamic scenario, with their Z-scores having very small variation under higher levels of perturbation in contrast to the other scoring functions.

### 5.6.3 Expand and Shrink

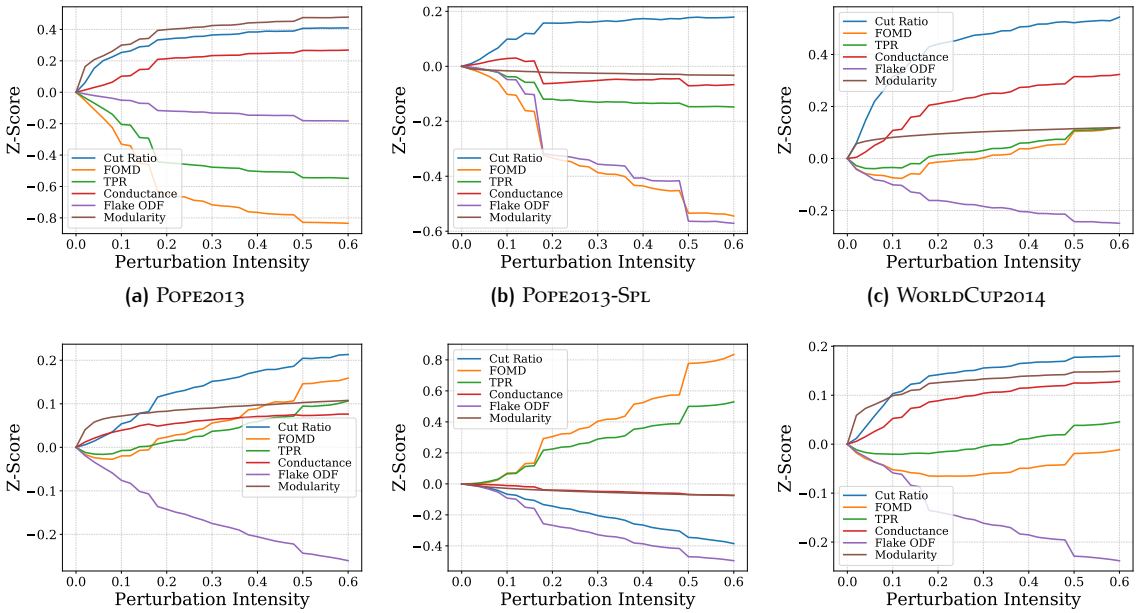
Lastly, the **Expand** and **Shrink** perturbations results seen in Figure 5.14 and Figure 5.15 also confirm TPR and FOMD as generally robust scores for Twitter functional communities in the dynamic scenario using user activity hotspots, specially the Shrink perturbation. Likewise the static




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

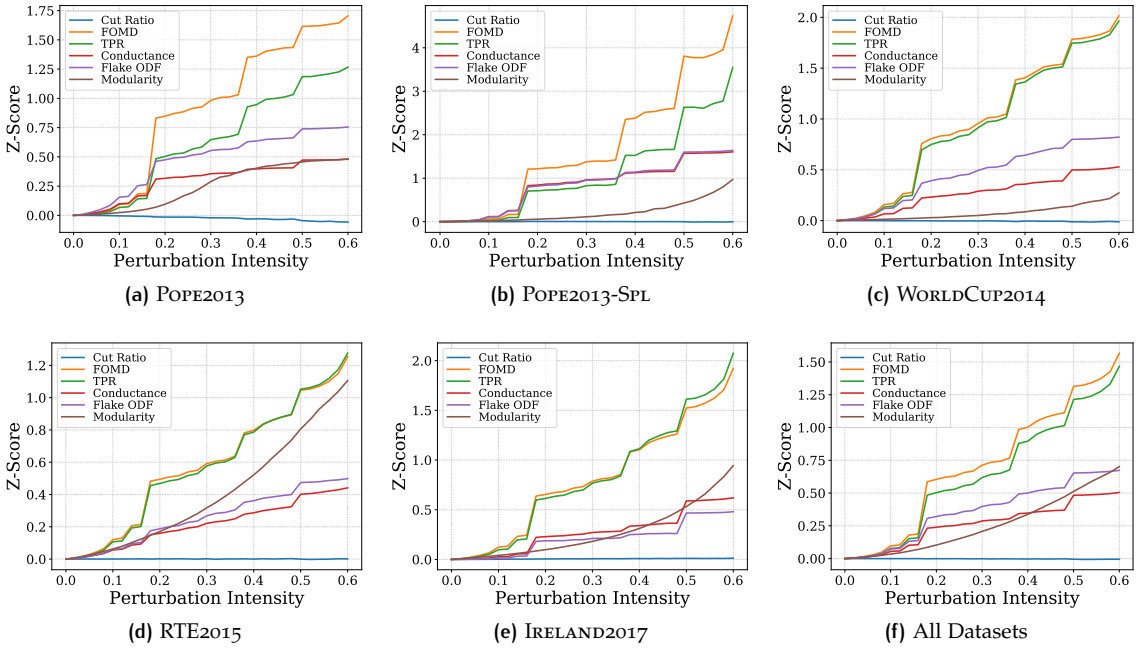
**Figure 5.13:** Z-scores of intensities for the Random perturbation strategy applied to all temporal sub-communities $H(C)$ based on user activity hotspots generated from communities $C$ for all community types in each Twitter ground-truth dataset. A combined plot is also presented.




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

**Figure 5.14:** Z-scores of intensities for the Expand perturbation strategy applied to all temporal sub-communities $H(C)$ based on user activity hotspots generated from communities $C$ for all community types in each Twitter ground-truth dataset. A combined plot is also presented.




**(a)** Pope2013 **(b)** Pope2013-Spl **(c)** WorldCup2014

**(d)** RTE2015 **(e)** Ireland2017 **(f)** All Datasets

**Figure 5.15:** Z-scores of intensities for the Shrink perturbation strategy applied to all temporal sub-communities $H(C)$ based on user activity hotspots generated from communities $C$ for all community types in each Twitter ground-truth dataset. A combined plot is also presented.

scenario, The Cut Ratio score is unable to handle communities that get smaller in Twitter, however its robustness improves when communities expand. The Modularity score performs more robustly than in the static case in the larger WorldCup2014, RTE2015 and Ireland2017 datasets for the Shrink strategy. However, for the other datasets it remains consistently well performing in small intensities for Expand and large for Shrink, but again degrades with larger expansions and smaller reductions, similarly to the static case. The resolution limit of Modularity [FB07] still produces a negative effect in the temporal sub-communities $H(C)$.

In this new experiment, TPR and FOMD from the internal connectivity family again proved to be robust community scoring functions for Twitter interaction streams in a dynamic setting, while Modularity and Cut Ratio proved weaker in the same context. Modularity, however, demonstrated to be more robust in the temporal sub-communities $H(C)$ due to their smaller sizes compared to the original communities $C$. Alternatively, Flake ODF and Conductance – in a lesser degree – from the mixed connectivity family are also reasonably robust choices for microblogging data in the dynamic scenario using user activity hotspots.

**Table 5.8:** Average absolute increment of Z-score between small ($p = 0.04$) and large ($p = 0.20$) community perturbations for the dynamic scenario. Largest differences (most robust and sensitive scores) are in bold. The previous best scores in the static scenario (Table 4.3) are marked with $\diamond$.

| Family | Score | N.Swap | Random | Expand | Shrink |
|---|---|---|---|---|---|
| External | Cut Ratio | 0.1032 | 0.0639 | 0.1072 | 0.0007 |
| Internal | FOMD | $\diamond$ **0.6098** | $\diamond$ **1.0362** | $\diamond$ 0.0361 | $\diamond$ **0.5902** |
| Internal | TPR | 0.4790 | 0.8623 | 0.0005 | 0.4919 |
| Mixed | Conductance | 0.3549 | 0.6627 | 0.0732 | 0.2298 |
| Mixed | Flake ODF | 0.2738 | 0.7174 | **0.1127** | 0.3099 |
| Net-Model | Modularity | 0.0591 | 0.0540 | 0.0525 | 0.0930 |

### 5.6.4 Detection Sensitivity

Finally, the sensitivity of the scoring functions in terms of small and large perturbations is re-investigated for the dynamic scenario presented in this chapter. For this experiment, again the change of Z-score between a small ($p = 0.04$) and a large ($p = 0.20$) perturbation is measured, giving preference to scoring functions that quickly degrade in presence of strong perturbations. The difference $Z(f, h, 0.20) - Z(f, h, 0.04)$ is averaged across all ground-truth temporal sub-communities $H(C)$ in all community type and datasets, and the results can be seen in Table 5.8. In these results, large differences indicate that the community scoring function is both robust and sensitive, and the previous bests from Section 4.3 are marked with $\diamond$.

In this new experiment, FOMD again remains as the most robust and sensitive scoring function for the dynamic scenario. Nevertheless, important differences can be highlighted with respect to the static case in Table 4.3. First, for the Node Swap perturbation FOMD has a positive difference (+0.1298) in the dynamic case than in the static case, however in the Random and Shrink strategies it has a negative difference ($-0.1289$ and $-0.1104$ respectively). This result indicates that, despite FOMD still being the preferred scoring function, it is also slightly less robust and sensitive in comparison to the static scenario. Moreover, Flake ODF and Cut Ratio surpass FOMD for the Expand perturbation in comparison with the static case, suggesting that in expanding communities the external connectivity is also necessary to be considered.

In general, the FOMD and TPR scores (internal connectivity family) still stand as the most robust and sensitive scores in this experiment for all the perturbation strategies but Expand – where the mixed and external connectivity scores are better – in the dynamic scenario. The Modularity score performs poorly under every perturbation strategy except Shrink, where only Cut Ratio is worse for microblogging data using user activity hotspots.

## 5.7 ACTIVITY HOTSPOTS DETECTION BIAS

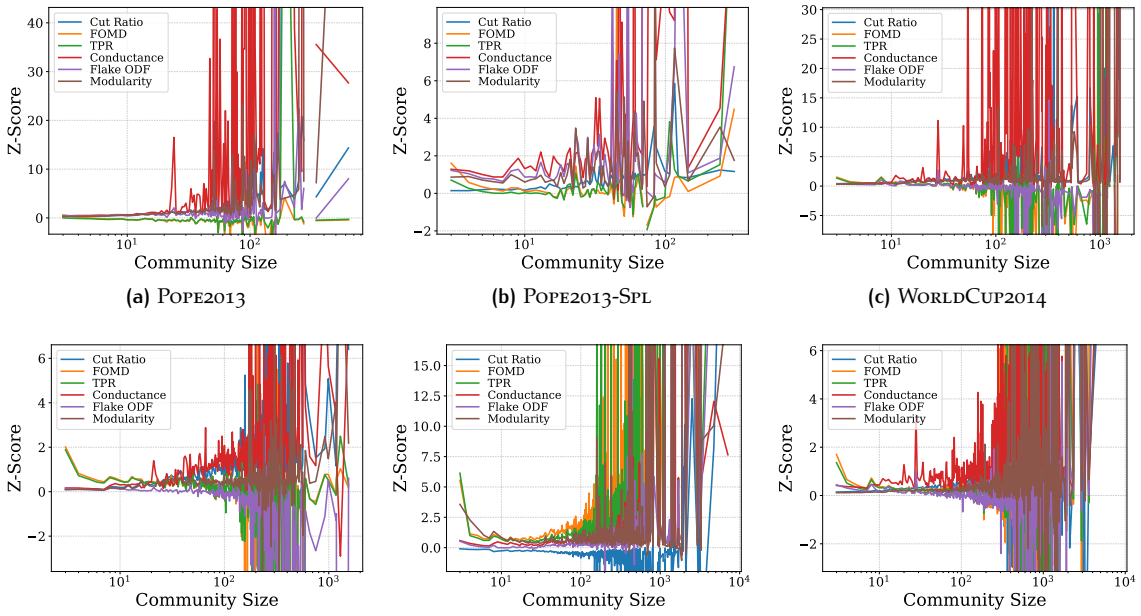
Similar to the results in Section 4.4, the observations for the Random perturbation strategy also revealed large differences in the robustness of the scoring functions. In Figure 5.13, the reported Z-scores for the combined datasets go up to 6.5 standard deviations from the mean, and in the case of the Pope2013-Spl dataset, as high as 15.0 standard deviations from the mean. The scoring functions might still be subject to a community size bias for the dynamic scenario using activity hotspots, where small communities are artificially over-scored. Therefore, the perturbation bias experiment is also re-investigated for the microblogging dynamic scenario.

The experiment is setup as follows. First, a relatively high ($p = 0.20$) constant perturbation intensity is chosen. Then, the changes in the Z-score as a function of the temporal sub-community sizes for the selected perturbation intensity is observed. Each Z-score is calculated with respect to all the temporal sub-communities with a given size. Because $p = 0.20$ represents a moderately strong intensity for all the investigated perturbation strategies, high values of Z-score and independent of the community size, i.e. constant, are expected if the scores are unbiased.

Figures 5.16, 5.17, 5.18 and 5.19 respectively show the new results for the Node Swap, Random, Shrink and Expand perturbation strategies under the chosen $p = 0.20$ intensity for all the temporal sub-communities in each ground-truth dataset, including a plot with all the data combined. As in the static case experiment, the results again contained very large Z-score values that subsumed the majority of the smaller values and the same outliers filtering strategy is adopted. the results are very similar to their static scenario counterparts in Section 4.5.

### 5.7.1 Node Swap

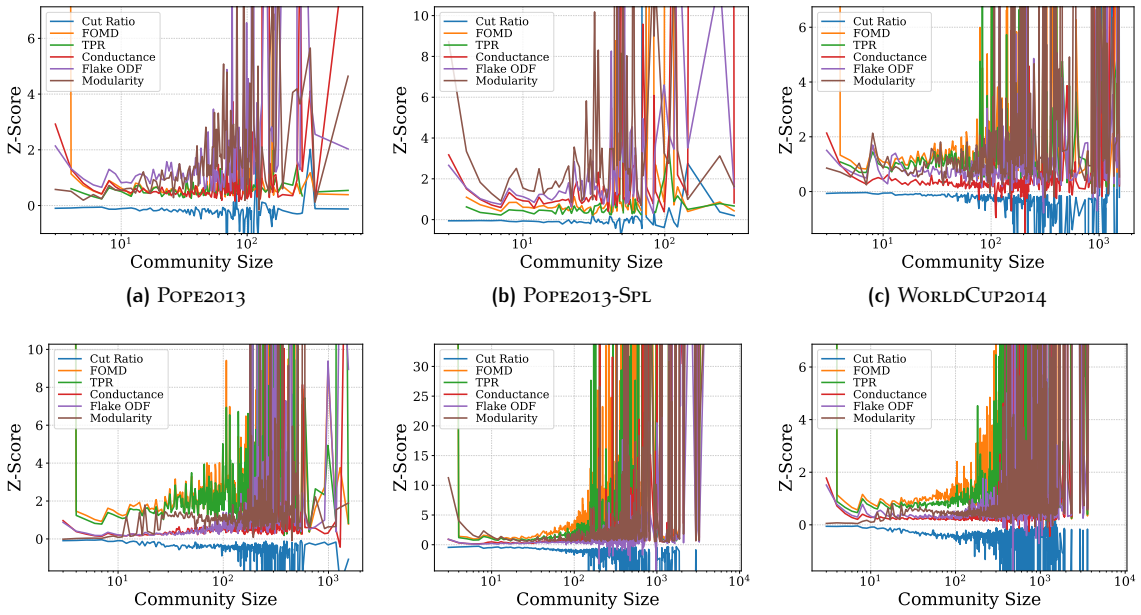
For the **Node Swap** perturbation in Figure 5.16, every scoring function in the experiment is not robust for small communities, e.g. sizes up to $\approx 10^{1.5}$ for the Pope2013, Pope2013-Spl and RTE2015 datasets, and sizes up to $\approx 10^{2}$ for the WorldCup2014 and Ireland2017 datasets. After these size limits, the values of Z-score are much higher. FOMD, TPR and Modularity are the exception to the above observation for the case of Pope2013-Spl, RTE2015 and Ireland2017, where they exhibit good robustness with smaller communities.




(a) Pope2013 (b) Pope2013-Spl (c) WorldCup2014

(d) RTE2015 (e) Ireland2017 (f) All Datasets

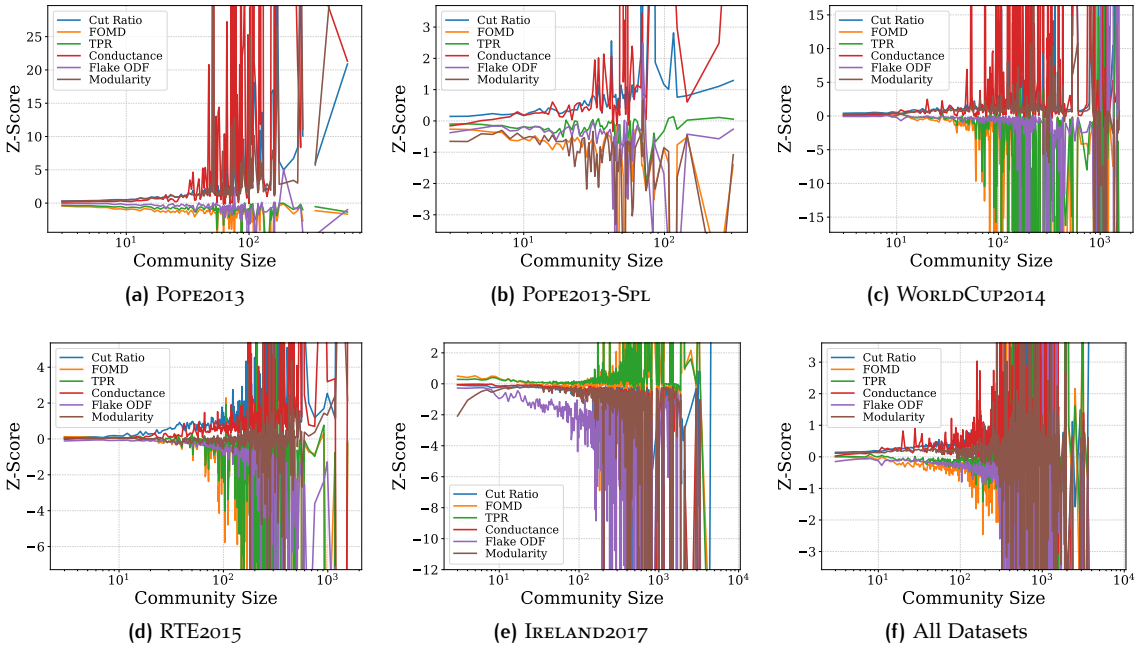
**Figure 5.16:** Z-scores as a function of community size for the Node Swap perturbation strategy applied to all temporal sub-communities $H(C)$ generated from communities $C$ for all community types for each Twitter ground-truth dataset. A combined plot is also presented.




(a) Pope2013 (b) Pope2013-Spl (c) WorldCup2014

(d) RTE2015 (e) Ireland2017 (f) All Datasets

**Figure 5.17:** Z-scores as a function of community size for the Random perturbation strategy applied to all temporal sub-communities $H(C)$ generated from communities $C$ for all community types for each Twitter ground-truth dataset. A combined plot is also presented.




(a) Pope2013 (b) Pope2013-Spl (c) WorldCup2014

(d) RTE2015 (e) Ireland2017 (f) All Datasets

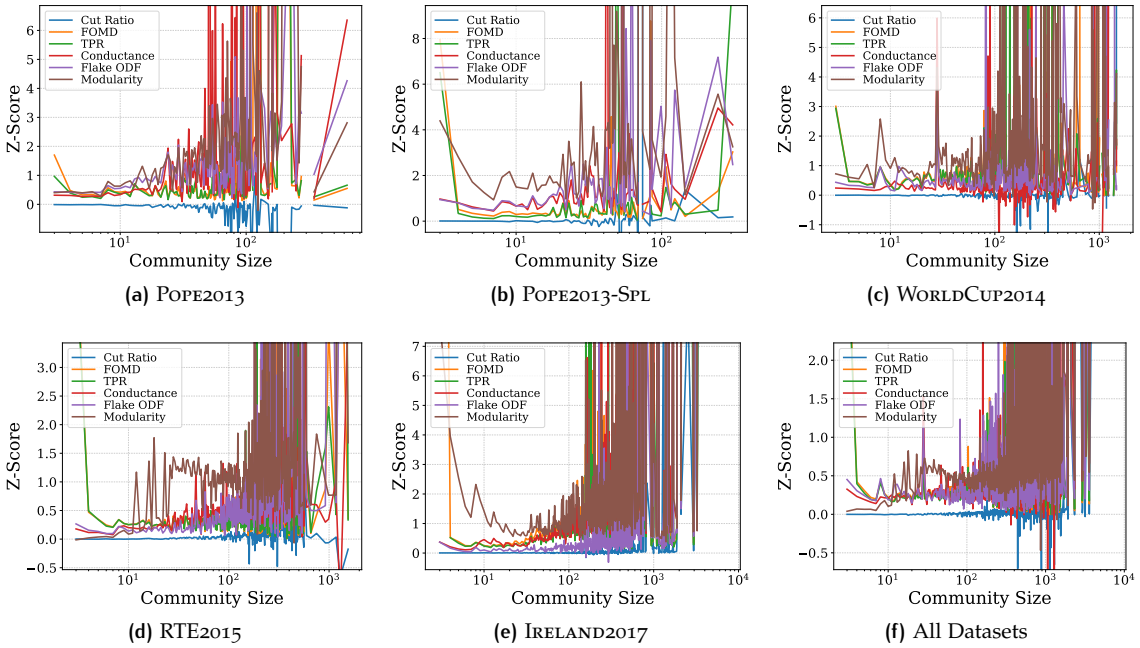
**Figure 5.18:** Z-scores as a function of community size for the Expand perturbation strategy applied to all temporal sub-communities $H(C)$ generated from communities $C$ for all community types for each Twitter ground-truth dataset. A combined plot is also presented.

### 5.7.2 Random

For the **Random** perturbation strategy in Figure 5.17, a very similar behaviour is again observed compared to the static scenario. In the dynamic case, again FOMD, TPR, Conductance and Flake ODF have more consistent robustness across community sizes. Cut Ratio remains stable but with Z-score values close to zero, suggesting that it is not able to distinguish perturbed and non-perturbed communities when the sizes are small enough, e.g. less than $\approx 10^2$ for the Pope2013 and Pope2013-Spl datasets.

### 5.7.3 Expand and Shrink

Lastly, for the **Expand** and **Shrink** perturbations seen in Figures 5.18 and 5.19, the results again reveal that the scoring functions have a bias for smaller communities, specially in the Expand strategy. Conductance and Flake ODF (mixed connectivity family) are the more robust in bigger ground-truth communities also for the dynamic scenario. On the other hand, for the Shrink perturbation, the Modularity scoring function is still prominently more robust on larger communities, again evidencing that its resolution limit also applies to dynamic microblogging data.




(a) POPE2013 (b) POPE2013-SPL (c) WORLDCUP2014

(d) RTE2015 (e) IRELAND2017 (f) All Datasets

**Figure 5.19:** Z-scores as a function of community size for the Shrink perturbation strategy applied to all temporal sub-communities $H(C)$ generated from communities $C$ for all community types for each Twitter ground-truth dataset. A combined plot is also presented.

In general, this experiment confirms that all the studied scoring functions also have an inherent bias towards small communities, i.e. produce artificially higher performance, for the dynamic scenario of microblogging social streams. In particular, some of the scores, e.g. Cut Ratio and Modularity, do not perform well when applied to communities smaller than $\approx 10^2$ users, no matter if considering a static or dynamic scenario. Nevertheless, the identified size bias does not render the scoring functions incapable of working in a dynamic microblogging data. Like in the static case, the scores belonging to the internal and mixed connectivity families proved the most robust and reliable for the analyst to consider, given that the temporal sub-communities under study are sufficiently large. Alternatively, the Conductance and Flake ODF scores are also good candidates for consideration in a lesser degree.

## 5.8 CHAPTER SUMMARY

In this chapter, the problem of further evaluating community detection in the context of microblogging services – represented by Twitter – in a dynamic scenario using user activity as a basis to generate temporal sub-communities was addressed.

First, an approach for extracting user activity from the experimental ground-truth communities was proposed. The method accumulates user activity in the underlying interactions network and applies a smoothing technique to minimise the effect of noisy data. Afterwards, a normalised activation threshold is proposed for the user activity that allows to identify starting and ending points in time of activity in the communities, called *activity hotspots*.

The same set of structural community scoring functions introduced in Chapter 4 were also evaluated using the constructed temporal sub-communities using the activity hotspots identified in microblogging streams, i.e. a dynamic scenario. This evaluation investigated the community detection goodness of the scoring functions and their robustness to a number of perturbation strategies. Furthermore, the sensitivity and bias of the scoring functions were also studied.

In the **dynamic scenario** of microblogging, activity hotspots in underlying user interactions networks from Twitter streams can be identified for ground-truth communities defined using social functions. From these hotspots, time-scoped functional sub-communities can be considered. Then, structural community definitions better align to these temporal functional sub-communities individually than to the whole original non-separated communities. As demonstrated in this chapter, the construction of activity hotspots from ground-truth functional communities further improves the ability of scoring functions based on internal and mixed connectivity such as TPR, FOMD and Conductance to discover communities in Twitter. Moreover, the Modularity score was shown to be less limited in this context.

# 6 PRACTICAL APPLICATIONS

In this chapter, real-world prototype practical applications developed in the context of this thesis are introduced as motivating examples of community detection for microblogging social media, represented by Twitter. Furthermore, a discussion of how the proposed user activity hotspots model in this thesis can improve their intended purpose is presented. The following main research question, proposed for this stage in Chapter 1, is addressed.

(RQ4) → *How can the findings from (RQ3) be integrated into real-world practical applications designed for different types of microblogging users, that utilise common community detection methods over microblogging data in their workflow?*

The example prototype applications are divided into two classes: (1) applications for **end-users**, aimed and designed for regular users of the social network itself, and (2) applications for **decision makers**, focused instead in supporting community owners and managers of the social network. For each of the above applications, their original purpose, functionalities, baseline approach for community detection and their inherent shortcomings are discussed.

An *end-user* is a common user that utilises Twitter and its social functionalities in a regular basis to get relevant content according to her interests. The prototypes for end-users include a form of community detection applied to live Twitter data for discovering user communities in near real-time. The application then provides these to the end-user in a curated manner for the users to better explore what is happening in Twitter around their interests.

Community owners/managers are individuals responsible for maintaining online communities related to their interests in a desirable state, e.g. active or relevant. In this thesis these users are termed the *decision makers*, after their ability to make decisions that potentially affect the development of their communities. Decision makers often require to monitor social networks under their supervision for potential problems and opportunities such as low activity, topic divergence, and user churn [KRC+11]. Then, according to this feedback, the decision maker might act accordingly to minimise any observed negative effect.

The *Whassappi!* mobile application prototype is first presented. *Whassappi!* was developed for the Galway Volvo Ocean Race event in 2012[1]. It consists of three main screens that showcase discovered communities and important users to follow related to the race event. Afterwards the *RTÉ Xplorer* web application prototype is introduced. *RTÉ Xplorer* was developed in conjunction with the national Irish television and radio broadcaster RTÉ (Raidió Teilifís Éireann)[2]. It consists of two main views, first a selection of television shows using interactive carousels and then a per-programme view with discovered topical communities and important users to follow related to each television programme being broadcasted. Both applications rely on the modularity scoring function implemented by the OSLOM method [LRR+11] for performing community discovery, and on the PageRank algorithm [PBM+99] for the Twitter user ranking functionality.

Lastly, for decision makers, The *DCV* (Dynamic Community Viewer) prototype is introduced. This application is a versatile tool designed for visualising the development of discovered user communities at different points in time. The *DCV* is capable of visualising evolutionary events such as birth, grow, shrink and death for dynamic user communities. This application includes community detection applied to live Twitter data for discovering user communities in near real-time, to later provide user community metrics intended for decision makers.

The main contribution of this chapter is the discussion of different motivation demonstrators for the proposed dynamic detection model for microblogging, classified according to the different audiences that this thesis is relevant for. All of the practical applications in this chapter have been accepted for publication (*Whassappi!* [HKW+12], *RTÉ Xplorer* [HBH+17], *DCV* [HH18b]) and were developed jointly as part of the research in this thesis.

## 6.1 THE WHASSAPPI! PROTOTYPE

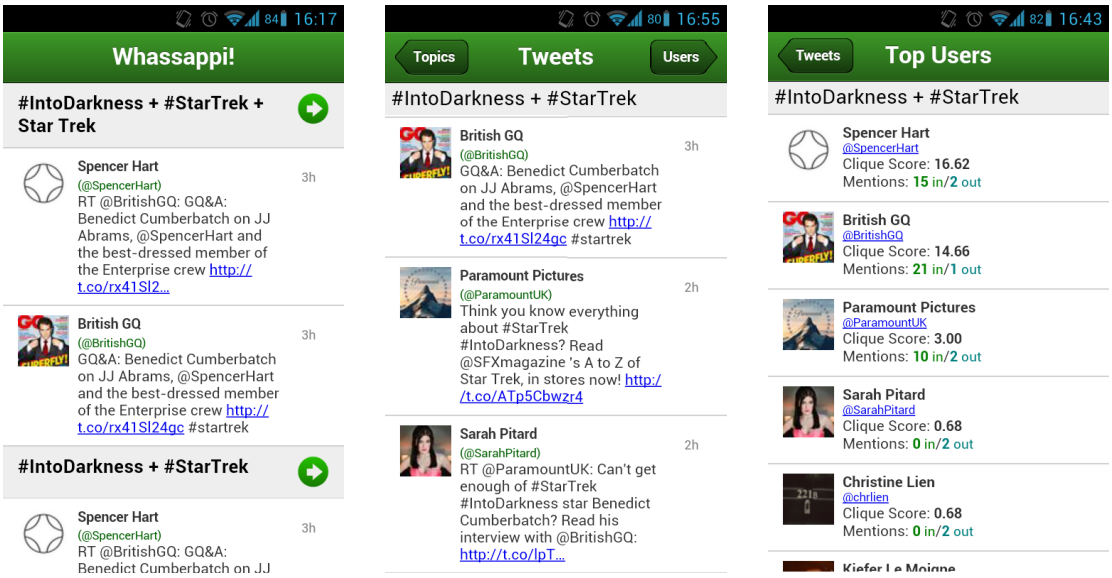
The first application, the *Whassappi!* prototype [HKW+12], will be introduced by quoting its official description[3]. It is redacted for its intended audience, i.e. Twitter end-users.

> "**Whassappi** (formerly known as: *Tweet Cliques*) is a small but smart app that tells you what's up and hot around public events (i.e. the Volvo Ocean Race) on Twitter. Like many apps. But Whassappi is **different**! It doesn't provide you a unified list with the trendy topics that every rag, tag and bobtail talks about. Whassappi shows you the things that groups of people around you, from small to large, are talking about.

1 https://en.wikipedia.org/wiki/2011%E2%80%9312_Volvo_Ocean_Race
2 http://www.rte.ie/about/
3 https://uimr.insight-centre.org/uimr/whassappi/

These groups, and all Tweets in the groups, are ranked by importance. Like this, you get a comprehensive view about everything happening around certain events at a glance. This makes Whassappi new and very different. Behind the scenes we use a smart mix of network analysis, topic monitoring, advanced labelling, social ranking and real-time processing - all in order to bring you a **new experience** in an intuitive and easy-to-use mobile interface."

To use *Whassappi!*, the user simply needs to open the installed application on their mobile device and the main screen is shown. This screen contains current **discovered communities** in Twitter related to a certain event of interest. From this point, the user can navigate the content using two additional screens showing details for particular communities and interesting users discovered for each of them. The three working screens of *Whassappi!* are shown in Figure 6.1.




**(a)** Live Communities (Main) **(b)** Community Tweets **(c)** Community Top Users

**Figure 6.1:** The Three Screens of the *Whassappi!* Prototype Application.

On the main screen (Figure 6.1a), *Whassappi!* displays a list of discovered topics and their associated user communities currently mined at the time of opening the application. The user communities in this screen are presented by listing the latest Tweets from the most central users in each group. If there are more communities in one topic, it shows more Tweets for that topic.

Next, if the user is interested on a particular topic and its communities, the green arrow in each topic allows her to move to the second screen (Figure 6.1b), where all the recent Tweets for that particular topic can be further explored. The Tweets in this list are ranked by the importance of the user that posted them, measured using the PageRank algorithm [PBM+99] over the underlying interactions network and the community participants.

Lastly, a third screen (Figure 6.1c) can be accessed from the Tweets list where the user can directly browse the ranked author users in the currently selected topic. To make this ranking more appealing to the general audience, the PageRank score is termed as the *Clique Score*.

All the content in the above views can be navigated freely, and also updated at any time by a drag-down gesture from the top of any screen, or by hitting the refresh button of the mobile device. This functionality allows the end-user to keep the application up to date with the current communities and important users being mined by the back-end system.

*Whassappi!* is not a replacement for a regular Twitter client. Instead the application is intended as a complementary tool designed to collate all those Tweets from the Twitter stream that are related to a particular setting. The application creates its own filtered stream from Twitter for a previously chosen event by an event manager, not by the end-users.

This prototype was designed and deployed in the context of the *Volvo Ocean Race 2012* world-wide event, particularly for its final leg in Galway, Ireland. The race began in Alicante, Spain, on 29 October 2011 and ended in Galway, Ireland, on 7 July 2012. For the occasion, Galway hosted a two-weeks free festival for the race finish, and it is estimated that "more then 820,000 people attended events during the first five days of the festival, with final attendance figures expected to be closer to the million mark"[4].

### 6.1.1 Limitations of the Application

In *Whassappi!*, the end-users have no control over the event being tracked or the global topic configured for the application. This aspect of the application is intended to be configured by the event owner and the system would then create its own filtered stream to process. This limitation adds an inherent bias from the community manager, however in practice this is mitigated using an adaptive approach for the listener component based on co-occurrence tables of listening terms [HKW+12]. The proposed method attempts to discard those terms that are not important (by frequency) to the main event being tracked.

The community detection approach used by this prototype is a simple application of the OSLOM algorithm [LRR+11], which is based on the modularity scoring function. Furthermore, the temporal aspect of the data is not taken into account by OSLOM directly. Instead, the system segments the input stream into configurable overlapping sliding windows, e.g. of one hour, that later are processed independently by the OSLOM algorithm. As reported in Chapter 4, the

4 http://www.thejournal.ie/volvo-ocean-race-brings-e100-million-into-galway-514188-Jul2012/

modularity scoring function is particularly weak in the case of Twitter data due to its resolution limit and sensitivity to the sparse characteristics of Twitter data.

## 6.2 THE RTÉ XPLORER PROTOTYPE

The *RTÉ Xplorer* prototype [BHH+16] is now introduced. This prototype is a web-based application designed to provide content adaptation and social awareness to end-users of RTÉ (Raidió Teilifís Éireann), the national provider of television and radio of Ireland. The semi-state company broadcasts its content online through their *RTÉ Player* service[5] and provides means to interact with its users using social media, such as Twitter and Facebook. RTÉ is interested in exploiting the potential of the knowledge immersed in their social media channels, and with this knowledge enhance their online services to further engage their viewers.

The overall goal of the *RTÉ Xplorer* prototype is to offer services for the RTÉ end-users that support them in exploring the RTÉ product catalogue and understand what is happening in social media in relation to RTÉ programming. In this manner, users could find interesting content faster and be encouraged to participate in social media communities discussing RTÉ content. For this, *RTÉ Xplorer* offers functionality based on both *Social Analytics* (via community detection) and *Information Adaptation*, using a simple recommendation engine based on the insights from the Social Analytics component. The prototype is a tangible representation of the above services that could be eventually integrated into the mainstream *RTÉ Player* service [BHH+16].

The *RTÉ Xplorer* prototype is composed of two views: (1) the *Exploration View* is the main landing page where the user can explore different programmes organised in sections adapted according to her viewing history and global social media activity, and (2) the *Programme View* is the detailed screen with all the information for a chosen video content, e.g. a programme episode, including live Tweets, discovered communities around the content and identified top users in those groups. An example of both views can be seen in Figures 6.2 and 6.3 respectively. The overall design of the prototype is inspired by the original *RTÉ Player* service.

This prototype is an extension to the simpler *Whassappi!* application functionality. User communities are not only discovered and aggregated towards specific RTÉ catalogue programmes, but also the co-occurrence of communities, member users and single Tweets are aggregated and used to feed a social-based, frequency-based item-item live recommendation engine.

5 https://www.rte.ie/player/

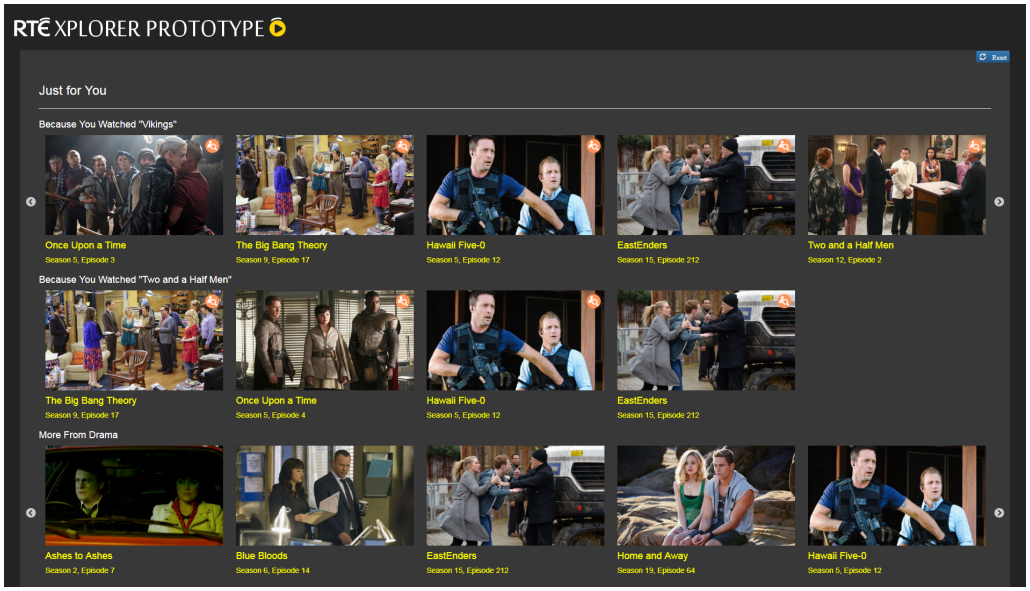



**Figure 6.2:** The Exploration View of the *RTÉ Xplorer* Prototype Application.

In order to offer services adapted to the unique characteristics of the RTÉ end-users, the system must cope with the following challenges: (1) lack of personal preference data such as ratings, (2) relatively small historical user session information, (3) dynamic inventory and limited life span of the recommendable items, (4) no direct integration of social media analytics, and (5) users need to be considered anonymous. The *RTÉ Xplorer* prototype addresses these challenges by identifying a key opportunity in using data immersed within social media, particularly in Twitter and its implicit communities, as a valuable resource that can be exploited to better understand user content preferences.

### 6.2.1 Limitations of the Application

The main research barrier for this prototype is the amount and quality of the data that the application has access to. In terms of user preference data related to programmes, RTÉ cannot freely share user profile data. In addition, given that RTÉ end-users are not required to register or sign-in to use the *RTÉ Player* service, many viewers do not have traceable user accounts. Even if RTÉ could share data freely, the amount of data still might not be enough to offer quality Information Adaptation services. On the other hand, in terms of programme-related data, the RTÉ programme catalogue is highly dynamic, i.e. the broadcaster adds and removes available programming on a daily basis. Traditional community detection approaches are difficult to apply over this fast changing behaviour. This is a major motivator for the research in this thesis.

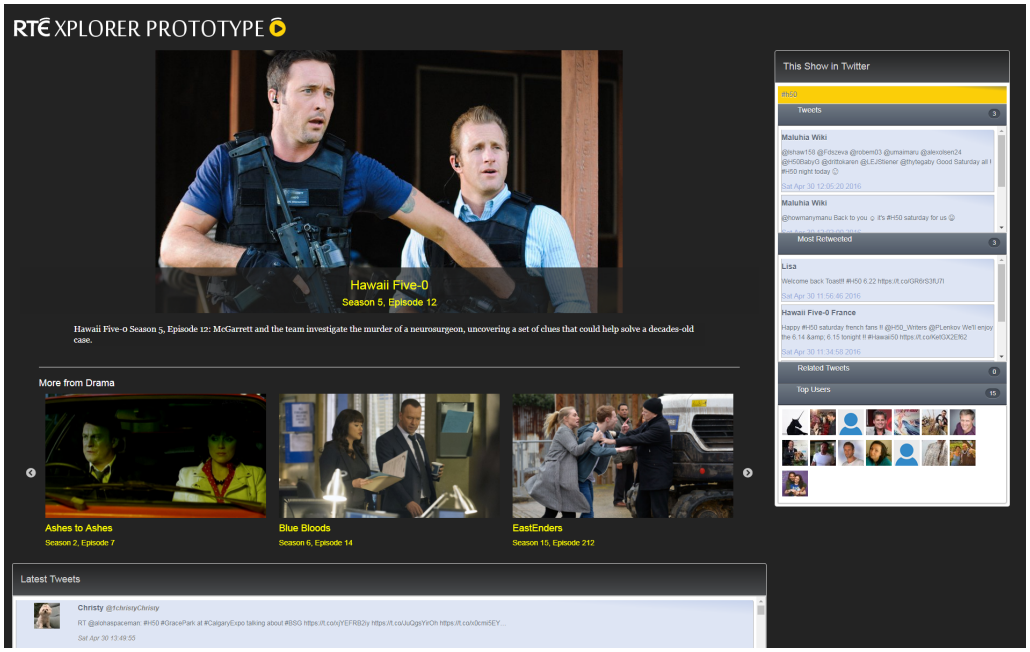



**Figure 6.3:** The Programme View of the *RTÉ Xplorer* Prototype Application.

Furthermore, this prototype uses the same community detection method in the *Whassappi!* prototype (the OSLOM method), therefore it also suffers from the same limitations. OSLOM is a non time-aware modularity scoring function with a weakness for the sparsity of Twitter data.

## 6.3 THE DYNAMIC COMMUNITY VISUALISATION (DCV) PROTOTYPE

One of the most interesting aspects of community detection for decision makers is the ability to visualise how their managed communities evolve in time. However, most community finding approaches are unable to generate such view directly and instead focus on being able to discover independent sets of communities statically at fixed periods. Fortunately, user community tracking algorithms exist [GDC10] that can analyse the discovered communities between consecutive time steps and identify connections between past and current configurations.

Using such tracking algorithms, a visualisation tool was developed [HH18b][6] for extracting dynamic tracking information for a set of discovered user communities to display their evolution based on different types of dynamic events that can occur: *birth*, *grow*, *shrink*, *intermittence*, *split*, *merge*, *death*, and *attribute change*, e.g. the community topic. The proposed workflow can be seen in Figure 6.4. First, the user chooses a community detection approach suitable for the dataset under study and extract sets of independent static communities $C_{t,i}$ for consecutive time

6 Available in: https://uimr.insight-centre.org/dcv/

steps t using any desired granularity, e.g. hourly or daily. With these static step communities, the user applies the TRACKER dynamic community tracking approach from Greene, Doyle, and Cunningham [GDC10] to obtain dynamic timelines that relate step communities to each other in time considering the defined evolutionary events. Finally, the user applies the TRACKER2VIS post-processing algorithm [HH18b] to compute additional evolutionary events, i.e. expansion, contraction and attribute change, and generate the complete visualisation information necessary for the interactive user interface.

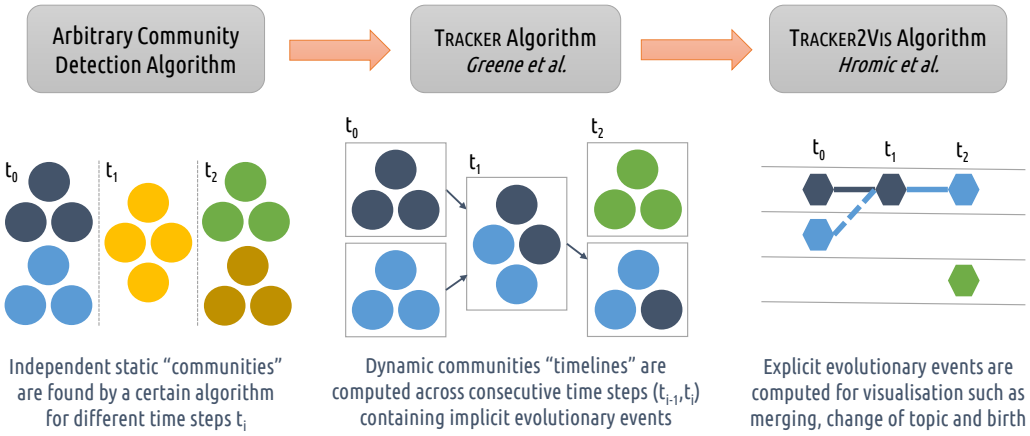



**Figure 6.4:** Usage workflow for visualising dynamic communities with the *DCV* Prototype Application.

The TRACKER method is based on the Jaccard similarity metric to compare "communities" across time steps, making this tracking approach highly convenient and flexible. Any group-like structure can be used in the *DCV* prototype and, despite the visualisation being designed towards community analytics, it is not strictly restricted only to this use case. In Figure 6.5, all of the supported evolutionary events in a synthetic set of user communities can be seen.

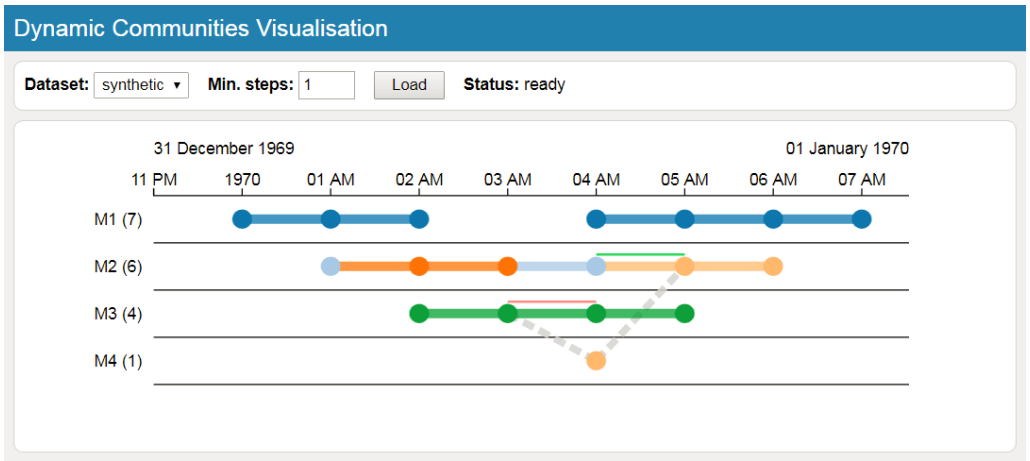



**Figure 6.5:** Example Dynamic Timeline of the *DCV* Prototype Application.

Four example dynamic community timelines can be seen named from $M_1$ to $M_4$. For each timeline $M_i$, static step communities $C_{i,t}$ exist with a particular topic label L in each time step t. Topic labels are represented with unique colours that can change in time. The extraction of topic labels is not restricted to any particular method. For example, in the *RTÉ Xplorer* prototype these topic labels can be mapped to RTÉ programmes associated to Tweets of the community, and for *Whassappi!*, the topic labels can be extracted using the most frequent hashtags in the community.

Furthermore, the *grow* and *shrink* is visualised using green and red over-lines respectively between two consecutive static community steps $C_{i,t}, C_{i,t+1}$. If the users of a static step community $C_{i,t}$ split, this is represented by a dashed line from the community to the new dynamic community these users migrated into. Likewise, if the members of a static step communities $C_{i,t}$ in a dynamic community $M_i$ join another existing dynamic community $M_j$ in a consecutive time step $C_{j,t+1}$, this is represented by a dashed line starting from the community that is joining to the merged community in the next time step. An example of these two events can be seen in the dynamic community $M_3$ at the second static step community $C_{3,2}$, where a portion of the users created a new dynamic community $M_4$, and later joined the dynamic community $M_2$.

### 6.3.1 Limitations of the Application

The *DCV* visualisation tool can provide a big picture of the evolution in time of a managed set of communities, however some limitations exist. While the visualisation is independent of the underlying community detection method used, the quality of the extracted dynamic events is subject to the quality of the detection approach and its robustness to the type of data being processed. An illustration of this using a real-world use case is in Figure 6.6, where the *DCV* prototype was applied to a Twitter dataset from the *RTÉ Xplorer* prototype system.

In this example, only dynamic communities with at least three static community steps are displayed. The community topic label (and colour) corresponds to the RTÉ programme that was referenced the most in each static community step. It can be observed that some dynamic communities such as $M_3$, $M_{27}$ and $M_{29}$ are steady and stable in time, while others such as $M_{12}$, $M_{14}$, $M_{21}$ and $M_{26}$ result in quite intermittent dynamic communities. After manual examination of these cases, it is revealed that there are very short-lived communities, as identified by the OSLOM method, however these communities tend to re-emerge later as new dynamic communities. This observation recurs in many other extracted dynamic communities in this dataset, and is not a desired behaviour as it can be misleading to the decision maker, who for example could

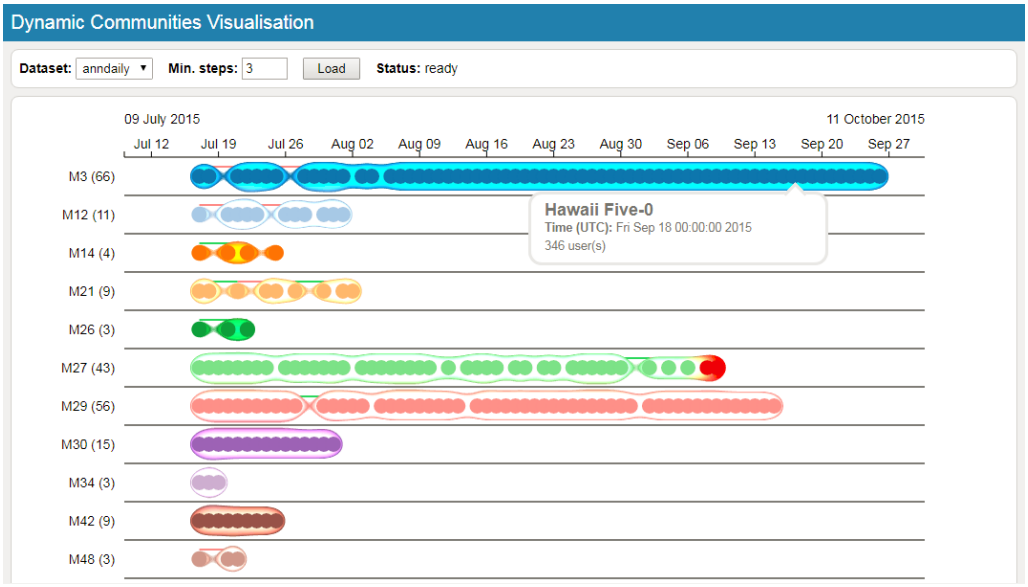



**Figure 6.6:** The *DCV* Prototype Application applied to RTÉ Twitter Data.

conclude that these short-lived communities are individual groups that are not performing well, while in fact they have been inaccurately detected as a single community.

Overall, further examination using the *DCV* visualisation tool suggests to the analyst that an alternative detection method might be necessary for the *RTÉ Xplorer* prototype. In particular, one that is more robust to the highly dynamic nature of the microblogging platform. While the original work from Greene, Doyle, and Cunningham reported good performance with traditional community detection methods on classic social media datasets, in the case of Twitter microblogging data the approach requires a method that is more robust to the highly dynamic nature of microblogging, as demonstrated in Chapter 4.

## 6.4 INTEGRATING USER ACTIVITY HOTSPOTS

The practical applications described in this chapter were originally designed to implement the modularity scoring function as a community detection approach for microblogging interactions networks. However, as investigated in Chapter 4, this approach is not adequate for the noisy and sparse data found in Twitter. Nevertheless, in Chapter 5, it was proposed that the performance of modularity and other scoring functions can be improved by first identifying user activity hotspots in the communities and then constructing temporal sub-communities.

To integrate the methods proposed in this thesis into practical applications of community detection for microblogging, the following strategy is proposed.

1. Select listening terms appropriate for the context of interest, e.g. events, topics or locations, and setup a stream capture from the desired microblogging service, i.e. Twitter.

2. Select a time quantisation $q$, e.g. hourly ($q = 3600$) or daily ($q = 86400$), and continuously construct a user interactions network from the stream as described in Section 3.3.

3. Select a window size $w_s = n \cdot q$, with $n$ sufficiently large, e.g. $n = 48$ for $q = 3600$, and collect microblogging data for at least $w_s$ time. If required, the network can be pruned using the stored temporal information, i.e. remove links older than $w_s$.

4. Using the guidelines in Section 5.2, estimate the time characteristic $\lambda$ for all the network edges, smooth the user activities and select an activation threshold $\alpha$, e.g. $\mathtt{alpha} = 0.2$, to activate ($> \alpha$) or deactivate ($< \alpha$) links in the network. Compute a structural community detection approach, e.g. modularity, to obtain user communities.

5. Update the interactions network with new data and repeat the process from step 3.

## 6.5 CHAPTER SUMMARY

Different practical applications of community detection for microblogging social networks were explored in this chapter. Two classes of users were identified: (1) end-users and (2) decision makers. For each class, two prototype applications were introduced that provide dynamic user communities to end-users and decision makers alike in a curated manner for exploration of what is happening in Twitter around their interests.

Two of the application prototypes (*Whassappi!* and *RTÉ Xplorer*) were originally designed to integrate the modularity scoring function implemented by the OSLOM method to enable community detection in their workflow. However, as demonstrated in Chapter 4, this scoring function is not ideal for microblogging social networks. Furthermore, these applications do not directly exploit the temporal information contained in the underlying user interactions networks. In Chapter 5 it was demonstrated that for better community detection performance in microblogging, it is necessary to implement a user activity hotspots extraction process and apply the desired community detection approach to the resulting temporal sub-communities. A general purpose strategy for this was presented in this chapter.

# 7 CONCLUSIONS

In this chapter, the general conclusions of the thesis and proposed future work are described. First, the conclusions of each chapter are collated and summarised to construct the final recommendations from this work in the context of microblogging community analytics. Afterwards, a summary of the contributions is discussed. Finally, the limitations and future directions proposed for the work carried on in this thesis is introduced and discussed.

## 7.1 SUMMARY OF THE THESIS

In this section, the conclusions from all the main research stages of the thesis are summarised. As introduced in Chapter 1, these four stages are: (1) the *preliminary* stage, (2) the *static scenario* stage, (3) the *dynamic scenario* stage, and (4) the *practical applications* stage. Each of them corresponds with a chapter in the thesis as described in the next sections.

### 7.1.1 Building Communities in Microblogging

In Chapter 3, the **preliminary stage** of the research was addressed. In this stage, the working definitions for functional and structural user communities were introduced for microblogging. Functional communities are defined as groups of users with a common and independent social *function*, e.g. fans of the same football team or people living in the same area, and structural communities are defined as groups of users that exclusively depend on their *connectivity in a network*, e.g. their average node degree or clustering coefficient. Furthermore, because the social functions in Twitter are explicitly-labelled by the users themselves, functional communities are deemed as ground-truth community data for the experiments in this thesis.

The task of evaluating community detection in the context of microblogging is defined as evaluating the quality of the alignment of structural community definitions to the defined functional

ground-truth communities. Therefore, a methodology for constructing ground-truth functional user communities from the Twitter microblogging social network was proposed and described, based on eight types of social functions. This methodology contemplates the building of both, functional communities and a user interactions network model for Twitter data. The user interactions network model constitutes the experimental data for the structural community definitions studied in this thesis. Five real-world Twitter datasets were captured and characterised, considering different world-wide events during a wide range of years using all the available capturing methods from Twitter at the time of this thesis.

#### *Research Questions and Sub-questions for the Preliminary Stage*

(RQ1) → *How can microblogging ground-truth and structural data for community detection be assembled and modelled in Twitter social streams?*

- (RQ1.1) → *For constructing ground-truth data, how can independent, explicitly user-labelled, functional communities be modelled in Twitter social streams?*

In microblogging, users create short messages that are rapidly spread among their followers and can contain a number of social functions that the users themselves attach to them. For example, user mentioning, user quoting, location or content tagging. These social functions can always be associated with a social object, e.g. the mentioned or quoted user, the particular location of the posting user or the topic associated to a content tag.

In Twitter, ground-truth functional communities are built 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 functional community $C_h = \{u_1, u_2, u_3\}$ is created. The following social functions from Twitter and their respective social objects were considered for building different ground-truth functional community types in this thesis: (1) hashtags, (2) mentions, (3) retweets, (4) quotes, (5) countries, (6) cities, (7) places and (8) URLs.

The participating users might not be fully aware of the social objects creating connections between them in the form of functional communities. It was demonstrated in the second stage of the research that such connection manifests through their live user interactions.

- (RQ1.2) → *For constructing structural data associated to the ground-truth in (RQ1.1), how can networks of user interactions be modelled in Twitter social streams?*

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 me-

dia files, e.g. pictures or videos. This special syntax, together with replying to posts, retweeting and quoting, can be used to form a network of interactions between users. The following types of Twitter interactions were considered for building a network of live user interactions from a stream of Tweets in this thesis: (1) mentions, (2) quotes, (3) replies and (4) retweets. For simplicity, in this thesis the interactions networks are considered as undirected.

Therefore, an undirected weighted network $G = (V, E, W)$ was proposed with a set of user vertices $u \in V$, interaction edges $e = (u_i, u_j) \in E$ and time-aware, typed, edge weights $w(e, t, type) \in W$. At every time $t$ a user $u_i$ interacts with another user $u_j$ using any of the defined interaction types in $T$, an edge $e = (u_i, u_j)$ is created in the network and the edge weight $w(e, t, type)$ is incremented by one. Furthermore, a quantisation function $Q(t, q) = \lfloor t/q \rceil \cdot q$, where $\lfloor \cdot \rceil$ is the nearest integer operator, is also applied to every time observation $t$. In this manner, edge weights $w(e, t, type) \in W$ can account for user activity per minute, hour, day or any other quantisation $q$ required by the researcher or analyst end-user.

Initially, building separate networks for each interaction type was considered, however a pairwise network overlap analysis in the experimental data revealed a low average network overlap of $\approx$2.39 %. It is then concluded that combining all the interaction types in a single network generates a richer structure overall while maintaining the simplicity of the model.

- (RQ1.3) → *What are the global properties, e.g. size and membership distributions, of the defined ground-truth functional communities in (RQ1.1) and (RQ1.2)?*

The general properties of the constructed ground-truth functional communities were investigated with the purpose of providing a better understanding of the ground-truth data used for the Twitter analysis. The following distributions of global properties were considered in this thesis: (1) community sizes, i.e. the number of users in the communities, (2) user memberships sizes, i.e. the number of communities that a user belongs to, (3) absolute community overlap sizes, i.e. the number of common users in a pair of overlapping communities, and (4) community ages, i.e. the length of time between the first user interaction and the last. To complement the global characterisation of community overlapping, the relative size of community overlaps was also investigated. This property measures how the ground-truth functional communities actually overlap: in nested structures or only for a small number of users.

It was observed that all of the distributions were heavily skewed. Most ground-truth functional communities are small, e.g. sizes between 1 and 10, however larger communities also exist, e.g. sizes $\approx 10^4$ and up to $10^6$. User memberships were found to be very sparse, with many users belonging to just a handful of communities ($<$10) and few users belonging to many communities,

e.g. generally $\approx 10^2$ and up to $\approx 10^4$. Community ages closely relate to the timespan of the captured datasets, e.g. many short-lived communities in the shorter datasets and many long-lived for the longer datasets. Community overlaps follow a power law and it was observed that many of the ground-truth functional communities do not overlap in general.

The properties investigated highlight the challenge of finding functional communities in the Twitter microblogging social network. Many of these communities are non-overlapping and built around very specific social objects, e.g. hashtags, mentioned users or specific locations. However, a measurable number of ground-truth communities evidenced user participation using multiple social objects at the same time. In conclusion, there is plenty of dynamic user activity in the ground-truth functional communities across all the constructed experimental datasets.

### 7.1.2 Characterising Communities in Microblogging

In Chapter 4, the **static scenario stage** of the research was addressed. In this stage, the problem of evaluating community detection in the context of microblogging services – represented by Twitter – was investigated. First, the structural properties of the constructed ground-truth functional communities in Chapter 3 were evaluated to investigate distinctive structural community characteristics in the data. Afterwards, a set of structural community scoring functions were evaluated using the functional ground-truth in a static scenario where no temporal information is considered from the microblogging streams. This evaluation investigated the community detection goodness of the scoring functions and their robustness to a number of perturbation strategies. Moreover, the sensitivity and bias of the scoring functions were also studied.

#### *Research Questions and Sub-questions for the Static Scenario Stage*

(RQ2) → *How do existing structural community definitions accommodate to microblogging ground-truth communities, including their robustness to random perturbations?*

- (RQ2.1) → *Do the defined ground-truth functional communities in (RQ1.1) evidence distinctive characteristics of structural communities, i.e. higher clustering coefficient, average degree, edge density and cohesiveness, in the associated networks of user interactions in (RQ1.2) in comparison to random groups with similar size and shortest-path distribution?*

To provide evidence of distinctive structural patterns in the network of user interactions, a comparative analysis of users in the ground-truth functional communities and randomly chosen connected nodes with the same path distribution was proposed. The following structural

properties were considered in this thesis: (1) clustering coefficient, (2) average degree, (3) edge density and (4) cohesiveness. In all the experimental datasets, a higher value for each structural property in the underlying interactions network was observed.

Functional communities built using the *mentions* social objects where found with the most prominent structural patterns in every dataset, suggesting that communities with a third person as functional object are easier to discover than other types of social objects. This result was similar for the *hashtags* social function. On the other hand, the location-based functional types in general contain few distinguishable communities due to low signal of Tweets containing useful location information. Nevertheless, the *countries* social object was found to be distinctive enough and proved to be the most suitable for building functional communities based on location.

- (RQ2.2) → *How well do state-of-the-art structural community definitions, e.g. based on triangle participation, conductance or modularity, align to the defined ground-truth functional communities in (RQ1.1) and (RQ1.2), including their robustness to random perturbations, e.g. member swapping, shrinking or expansion?*

Four families of structural scoring functions were considered in this thesis: (1) internal connectivity, (2) external connectivity, (3) mixed connectivity and (4) based on network models. Furthermore, thirteen scoring functions from these families were initially considered, and after a correlation analysis, six representatives were selected for further investigation in the thesis: (1) Fraction Over Median Degree (FOMD), (2) Triangle Participation Ratio (TPR), (3) Cut Ratio, (4) Conductance, (5) Flake Out Degree Fraction (Flake-ODF) and (6) Modularity.

In the static scenario, where temporal data is not considered, the scoring functions based on internal and mixed connectivity such as the TPR, FOMD and Conductance work the best for Twitter social networks, demonstrating to be robust and sensitive. Community detection approaches based on these definitions should be able to discover structural communities aligned to social functions in microblogging user interaction networks. On the other hand, the popular Modularity network model was found to be limited and unsuitable in this same context due to the sparse and noisy characteristics of microblogging data.

Furthermore, an inherent bias depending on the community size was identified in the evaluated scoring functions. Nevertheless, it was also demonstrated that once an adequate size of communities is considered, the scoring functions are robust enough for a researcher or analyst end-user to obtain reasonable results in the context of a static study.

### 7.1.3 Temporal Community Detection in Microblogging

In Chapter 5, the **dynamic scenario stage** of the research was addressed. In this stage, the problem of improving community detection in the context of microblogging services – represented by Twitter – in a dynamic scenario using user activity as a basis to generate temporal sub-communities was investigated. First, an approach for extracting user activity from the ground-truth communities was proposed. The method observes user activity in the underlying interactions network using time quantisation and applies an exponential decay smoothing technique to minimise the effect of noisy data. Afterwards, a normalised activation threshold was proposed for identifying starting and ending points in time of prominent user activity in the communities. These active ranges in the community are named *activity hotspots*.

The constructed temporal sub-communities based on activity hotspots were evaluated using the same set of structural community scoring functions introduced in Chapter 4. This evaluation investigated the improvement in community detection goodness of the scoring functions and their robustness to a number of perturbation strategies. Furthermore, the sensitivity and bias of the scoring functions were also studied again for this dynamic scenario.

#### *Research Questions and Sub-questions for the Dynamic Scenario Stage*

(RQ3) → *How can activity hotspots based on the dynamic user activity in time be identified in the defined ground-truth communities to improve community detection?*

- (RQ3.1) → *What are the temporal characteristics, for instance the user activity distributions, of the defined ground-truth functional communities in (RQ1.1) and (RQ1.2)?*

The time dimension is a fundamental aspect in microblogging social data due to its fast-pacing and sparsity as discussed in Chapter 4. Therefore, the dynamic scenario of microblogging where user communities are modelled considering their temporal properties was also investigated. It was observed that, given a period of time that is long enough, user interactions network data captured from microblogging social media tends to become increasingly complex and, in turn, discovering user communities in these networks becomes more difficult.

However, as discussed in Chapter 2, many community detection approaches do not consider any temporal information in their network representations. Moreover, if network data is captured for a long enough period of time, eventually some links become much more active than others. It was observed that in real-world social networks, and specially in microblogging, the user activity between nodes is not evenly distributed.

- (RQ3.2) → *Using the dynamic user activity in time as a basis, how can activity* hotspots *be identified in the defined ground-truth functional communities in (RQ1.1) and (RQ1.2) to be used for further identifying time-scoped sub-communities?*

The ground-truth functional communities from the static scenario can be divided using temporal windows according to the user activity recorded in the interactions network, forming temporal sub-communities, i.e. *activity hotspots*, within the static communities. Low user activity can lead to links being discarded or even larger parts of the static community to completely dissolve. Moreover, the same users can participate in different structures at different times. To discover these user activity hotspots, the accumulated user activity stored in the user interactions network was extracted for each ground-truth functional community. Afterwards, a predefined activity threshold was proposed for deciding the time boundaries of the temporal sub-communities. Twitter activity data was found to be noisy, therefore an exponential decay smoothing function was adopted that activates with activity and decays in time during inactive periods based on a time characteristic parameter. This smoothing function was shown to enable smoothed user activity measurements for a continuous range of time, using any granularity desired. To obtain a suitable activity threshold parameter, two quantitative metrics for activity hotspots that measure their basic characteristics for any candidate threshold parameter were proposed in the thesis: (1) the average number of hotspots per community (HpC), and (2) the average number of users per hotspot (UpH). The metric that exhibits the higher statistical coefficient of variation is selected as criterion to be maximised by the candidate threshold.

It was found in all the experimental datasets that no more than 30 % of the user activity per hour is required to activate the formation of reasonable hotspots in the ground-truth communities. In addition, the set of goodness metrics from Chapter 4 was investigated on the generated temporal sub-communities. Evidence was found that using activity hotspots enables communities to be more clustered and cohesive and therefore improve their chances of being discovered by structural approaches. Furthermore, it was found that the dominant criterion is HpC, suggesting that the average number of hotspots per community is preferred over the average amount of users in them for constructing adequate user activity hotspots for the dynamic scenario.

- (RQ3.3) → *Considering the identified time-scoped sub-communities based on user activity hotspots defined in (RQ3.2), how well do the state-of-the-art structural community definitions investigated in (RQ2.2) now align to these sub-communities in comparison to the ground-truth functional communities in the static scenario, i.e. without considering their user activity context?*

In this thesis it is suggested that temporal sub-communities constructed using user activity hotspots are of better structural quality than ground-truth communities where the dynamics are not taken in consideration. To provide preliminary evidence for this, the structural properties of the temporal sub-communities were studied using the same experimental approach described for (RQ2.1). Like in the original experiment, distinguishable structural patterns were observed in the activity hotspots in comparison to randomly chosen nodes with similar shortest path distribution. In general, all the structural properties for all the considered community types in all of the experimental datasets evidenced improvement over the base static scenario case. For instance, the location-based community types that previously did not exhibit distinguishable structural patterns in the static scenario, became discernible in the dynamic scenario.

The community goodness metrics defined in (RQ2.2) applied to temporal sub-communities based on activity hotspots were also investigated. In this thesis it was demonstrated that, when considering user activity for the identification of activity hotspots, the resulting temporal sub-communities are of better quality than user communities in a static setting. In comparison to the static scenario, all of the scoring functions are able to improve their ranked alignment to the goodness metrics at different degrees when considering the proposed activity hotspots. In general, the internal and mixed connectivity family of scores are the ones benefiting the most from using user activity hotspots to identify temporal sub-communities in the ground-truth. In a lesser degree, the other families also can benefit in certain cases. The activity hotspots were able to improve the performance of structural scoring functions in terms of Cohesiveness, Density and Clustering Coefficient, but not Separability, which remained mostly unaffected. Overall, the community goodness experiments carried in Chapter 5 suggest that, to identify more clustered, dense and cohesive communities in Twitter in a time-aware context, FOMD and TPR are the better choices for structural scoring functions. If the researcher wishes for dense but more separated communities, then Conductance or Cut Ratio should be considered.

The robustness and sensitivity of the scoring functions in the context of the temporal sub-communities generated using the proposed activity hotspots was also investigated. In general, the FOMD and TPR scores – of the internal connectivity family – still remained as the most robust and sensitive scores for most of the perturbation strategies under study for the dynamic scenario. The Modularity score performed poorly under every perturbation strategy except Shrink, where only Cut Ratio was worse in the experimental microblogging data.

Finally, in the dynamic scenario using activity hotspots, evidence of the scoring functions still being subject to a bias was found. In general, the same bias experiment from (RQ2.2) confirmed that all the studied scoring functions also have an inherent bias over the size of the communities

that they are measuring for the dynamic scenario of microblogging social streams. In particular, some of the scoring functions were found not strong when applied to communities smaller than $\approx 10^2$ users, regardless of a static or dynamic context. Nonetheless, the bias found did not render the scoring functions incapable of working in a dynamic microblogging data. Similar to the static case, the scores belonging to the internal and mixed connectivity families proved the most robust and reliable for the research to consider, given that the temporal sub-communities under study are sufficiently large. As an alternative, the Conductance and Flake ODF scores were also found as good candidates for consideration in a lesser degree.

### 7.1.4 Practical Applications

In Chapter 6, the **practical applications stage** of the research was addressed. In this stage, different demonstrator applications for the proposed dynamic model for microblogging communities were presented in this thesis, classified according to the different audiences they are aimed for.

The prototype applications were divided into two classes: (1) applications for **end-users**, aimed and designed for regular users of the social network itself, and (2) applications for **decision makers**, focused instead in supporting community owners and managers of the social network. For each of the applications introduced, their purpose, functionalities, baseline approach for community detection and their inherent shortcomings were discussed.

#### *Research Questions for the Practical Applications Stage*

(RQ4) → *How can the findings from (RQ3) be integrated into real-world practical applications designed for different types of microblogging users, that utilise common community detection methods over microblogging data in their workflow?*

All the prototypes discussed in this thesis include some form of community detection applied to live Twitter data for discovering user communities in near real-time. The applications provide dynamic user communities to end-users and decision makers alike in a curated manner for exploration of what is happening in Twitter around their interests.

Two practical applications involving community detection were proposed for end-users of microblogging social networks. First a mobile application prototype developed for the Galway Volvo Ocean Race event in 2012 was described. This mobile application consists of three main screens that showcase discovered topical communities and important users to follow related to the race event. Second, a web application prototype developed in conjunction with the national

Irish television and radio broadcaster RTÉ (Raidió Teilifís Éireann) was introduced. This web application consists of two main views, first a selection of television shows using interactive carousels and then a per-programme view with discovered topical communities and important users to follow related to each television programme being broadcasted. Both of these application prototypes were originally designed to integrate the modularity scoring function implemented by the OSLOM method [LRR+11] for performing community detection.

For the case of decision maker users, another two practical applications also involving community detection in microblogging were proposed. Decision makers often require to monitor the user communities under their supervision for potential problems and opportunities such as low activity, topic divergence, and user churn. Afterwards, according to this monitoring, they might act accordingly to minimise an observed negative effect. Therefore, prototype applications were proposed to provide instruments to visualise and measure different key aspects of the user communities under management. First, a dynamic community viewer prototype was introduced. It is a versatile tool designed for visualising the development of discovered user communities at different points in time. The prototype can visualise evolutionary events such as birth, grow, shrink and death for user communities. Afterwards, a third-party web application designed for managing enterprise-oriented social networks was introduced, where community detection and microblogging support was added for this thesis. The application can monitor a rich set of characteristics of online user forums, such as roles composition, tag-clouds and implicit community relationships, all of this using freely configurable time windows.

## 7.2 SUMMARY OF RESEARCH OUTCOMES

From all the research questions presented in the main chapters of this thesis, a set of fundamental research outcomes were stated in Chapter 1 that now are summarised and concluded below.

- Ground-truth communities defined using social functions have distinctive structural patterns in a network of live user interactions in microblogging. From: (RQ2.1).
- Microblogging users switch topics quickly and do not participate in steady and long-lived communities, rendering conventional structural community discovery approaches based on static and dense networks less effective in microblogging. From: (RQ1.3) and (RQ2.2).

Research evidence for these two expected outcomes was discussed in Chapters 3 and 4. The structural properties of the built ground-truth functional communities were evaluated and dis-

tinctive structural community characteristics were observed in the data. Furthermore, structural community scoring functions were investigated in a static scenario and an evaluation of the community detection goodness of these functions and their robustness to a number of perturbation strategies was discussed. The sensitivity and bias of the scoring functions were also studied.

- Activity hotspots in underlying user interactions networks from microblogging can be identified for ground-truth communities defined using social functions. From these hotspots, time-scoped functional sub-communities can be considered. Then, structural community definitions better align to these temporal functional sub-communities individually than to the whole original non-separated communities. From: (RQ3.1), (RQ3.2) and (RQ3.3).

Evidence for this expected outcome was provided in Chapter 5. After investigating the static scenario of microblogging, temporal sub-communities based on activity hotspots were proposed, constructed and evaluated using the same set of structural community scoring functions from Chapter 4. In the research experiments, a measurable improvement was found in the community detection goodness of the scores and their robustness to a number of perturbation strategies.

In general, all the research carried in the thesis leading to the above outcomes allow to better understand of how functional user communities can be uncovered from live streams of user interactions in microblogging social networks using time-aware methods.

## 7.3 SUMMARY OF CONTRIBUTIONS

Addressing the research questions of this work provided a number of contributions to the research field of community detection for microblogging, including dynamic community detection, visualisation and evaluation. The identified contributions of this thesis are below.

1. From (RQ1.1) and (RQ1.2), a methodology for building ground-truth functional communities from microblogging live user interactions. In particular for stream-based Twitter datasets.

2. From (RQ1.3), (RQ2.1) (RQ2.2), (RQ3.1) and (RQ3.3), an in-depth characterisation, understanding and evaluation of global and structural properties for functional communities in microblogging social media, for both the static and dynamic scenarios.

3. From (RQ2.2) and (RQ3.3), a set of recommendations on community detection algorithms based on data-driven evaluation of Twitter user interactions networks.

4. From (RQ3.2), a strategy for the identification of activity hotspots in functional communities in microblogging based on the network of user interactions, that improves the performance of existing community detection algorithms designed for static data.

5. From (RQ4), a set of four demonstration applications – two for end-users and two for decision makers – designed for microblogging social media, including a community detection system and a dynamic communities visualisation tool.

6. An open source implementation of the analytical framework developed for this study[1].

### 7.3.1 Publications

The following publications emanated during the development of this thesis.

- H. Hromic and C. Hayes. "Characterising and Evaluating Online Communities from Live Microblogging User Interactions". In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. Aug. 2018, pp. 21–24

- H. Hromic and C. Hayes. "Visualising the Evolution of Dynamic Communities in Social Networks using Timelines". In: *3nd ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data (AALTD)*. The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD). Sept. 2018

- H. Hromic, A. Barraza-Urbina, C. Hayes, et al. "Mining TV Twitter Networks for Adaptive Content Navigation and Community Awareness". In: *Expert Update*. AI-2016 Thirty-sixth SGAI International Conference on Artificial Intelligence 17.1 (2017)

- H. Hromic, N. Prangnawarat, I. Hulpuş, et al. "Graph-Based Methods for Clustering Topics of Interest in Twitter". In: *Engineering the Web in the Big Data Era*. Ed. by P. Cimiano, F. Frasincar, G.-J. Houben, et al. Lecture Notes in Computer Science. Springer International Publishing, 2015, pp. 701–704. ISBN: 978-3-319-19890-3

- H. Hromic and C. Hayes. "Constructing Twitter Datasets using Signals for Event Detection Evaluation". In: *Synergies of Case-Based Reasoning and Data Mining Workshop*. 22nd International Conference on Case-Based Reasoning. Sept. 29, 2014

- H. Hromic, M. Karnstedt, M. Wang, et al. "Event Panning in a Stream of Big Data". In: *LWA Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML)*. 2012

1 Online public repository: https://github.com/hhromic/dyncom-analytics

## 7.4 LIMITATIONS AND FUTURE DIRECTIONS

The research carried in this thesis has a number of limitations that must be considered by the reader. First and foremost, the experimental datasets were constructed in Chapter 3 based on the Twitter platform as representative of microblogging social media. While Twitter is a suitable representative due to its wide array of social functionalities for microblogging (refer to Section 2.3.1 and Appendix A), the manner in which the users utilise these functionalities may be different between microblogging platforms. Moreover, other external characteristic such as the cultural context of the users, for instance the Chinese platform Sina Weibo, might play an important role on how online microblogging communities form and develop statically and dynamically.

Another identified limitation is the difficulty of reliably capturing Twitter data, as discussed in Section 3.4. In particular, it is extremely hard to get abundant and unbiased social data from this platform due its high volume and price for accessing. For this reason, the most common method for capturing Twitter data consists on filtering the global stream using keyword terms, users to follow or geo-bounding boxes. This approach conveys an inherent bias on how the filter is configured. In this thesis, a best effort was made to minimise the potential listening bias, including the construction of a topic-independent dataset (i.e. IRELAND2017). However, trending topics and popular users might still be captured more prominently than less used social objects.

Lastly, in Chapters 4 and 5, an inherent bias towards the size of ground-truth functional communities was also identified for the structural scoring functions. If the ground-truth communities are too small, e.g. less than 100 members, then it becomes more difficult for the selected scoring functions to distinguish community structures in the network of user interactions. As studied in Chapter 3, a large number of the constructed ground-truth communities in the experimental datasets are in fact small. While the results in the thesis are still considered significant in the scope of the proposed research (as discussed in the respective chapters), future steps into further studying community detection for microblogging should take this bias in consideration.

### 7.4.1 Future Directions

More research is required to further better understand the nature of microblogging social networks and the community detection task for them. For example, native hashtags used as social function for topics were considered in this thesis, however other models can be used instead for the same social function, e.g. named entities, bag-of-words or TF-IDF. Moreover, the construc-

tion and exploitation of user activity hotspots can be further improved. For example, the time characteristic parameter ($\lambda$) and the activation threshold ($\alpha$) could be learned from empirical data using machine learning techniques. Furthermore, the current findings in this thesis such as the activity threshold can potentially be applied to the design of an automatic windowing approach for real-time community detection using stream processing methods.

In the practical applications presented in this thesis, it is important to note that the end-users have no control over what is being tracked in microblogging or what is the global topic for the application. This aspect was intended to be configured by the decision maker, i.e. community owner, and the system would then create its own filtered stream to process. This limitation adds an inherent bias from the decision maker, however in practice this is mitigated using a simple adaptive approach for the listener component based on co-occurrence tables of listening terms [HKW+12]. This method attempts to discard those terms that are not important (by frequency) to the social context being tracked in microblogging. While this approach was sufficient as a prototype, there is still room for improvement. For example, a more advanced adaptive system could be constructed that not only considers active co-occurrence but also historic terms.

The community detection approach originally used by all the prototype applications is an application of the OSLOM algorithm [LRR+11], which is based on the Modularity scoring function. As reported in Chapter 4, this scoring function is particularly weak in the case of microblogging data due to its resolution limit and sensitivity to the sparse characteristics of this social media. In addition, the temporal aspects of the data are not taken into account by OSLOM directly.

In general, the main research barrier for community detection applications is the amount and quality of the data that the application has access to in its social domain. Traditional community detection approaches are difficult to apply over the fast changing behaviour of microblogging, which is reflected on the user activity.

# Appendices

# A MICROBLOGGING SERVICES

In this appendix, a number of microblogging services and their history are reviewed in brief to further establish the importance of this platform for this thesis in the context of social media.

## A.1 HISTORY OF MICROBLOGGING

The term *microblog* did not originate immediately. In April 2005, the term *tumbleblog* was coined by Jonathan Gillette[1] in a discussion regarding blogging activities[2], where he notes how they were evolving. His description closely resembles key microblogging characteristics.

> "Blogging has mutated into simpler forms (specifically, link- and mob- and aud- and vid- variant), but I don't think I've seen a blog like Chris Neukirchen's [sic] Anarchaia, which fudges together a bunch of disparate forms of citation (links, quotes, flickrings) into a very long and narrow and distracted tumblelog."

Later that year in October, Jason Kottke[3] described tumblelogs more precisely[4].

> "A tumblelog is a quick and dirty stream of consciousness, a bit like a remaindered links style linklog but with more than just links. They remind me of an older style of blogging, back when people did sites by hand, before Movable Type made post titles all but mandatory, blog entries turned into short magazine articles, and posts belonged to a conversation distributed throughout the entire blogosphere. Robot Wisdom and Bifurcated Rivets are two older style weblogs that feel very much like these tumblelogs with minimal commentary, little cross-blog chatter, the barest whiff of a finished published work, almost pure editing...really just a way to quickly publish the "stuff" that you run across every day on the web."

1 Gillette is a writer, cartoonist, artist, programmer and a key figure in the Ruby programming language community.
2 https://viewsourcecode.org/why/redhanded/inspect/tumbleloggingAssortedLarvae.html
3 Kottke is an American blogger, winner of a Lifetime Achievement Award for blogging. Since 2013, his blog is ranked #66 overall and #20 in Science on the Technorati Top 100 ranking.
4 https://www.kottke.org/05/10/tumblelogs

It can be noted how Kottke remarks that people started to evolve from long and careful writings in blogs to more straightforward and mundane content into the web. There was a need to publish shorter but more often than to write extensively for longer periods. Web users started to modernise and instead of preparing long narrations of their lives in daily, weekly or monthly doses, found it more practical to just report their experiences shortly in real-time, i.e. hourly.

Since 2006, a number of services providing the above distinctive functionalities have been launched that helped establish the modern *microblog* term. Some of them have seen success while others failed to stay relevant. Moreover, these successful platforms also have had varied levels of adoption by the public. Twitter is perhaps one of the most widely known microblogging OSN in the world, with more than 330 million monthly active users as of December 2017[5], generating an average of 500 million *Tweets* (short messages) per day.

To showcase the rich history of microblogging social media, seventeen representative services will be introduced below in chronological order. All these platforms have influenced the development of microblogs at different degrees during the recent years.

**JAIKU** is one one of the earliest known services offering microblogging and life-streaming functionalities[6]. Jaiku was founded in February 2006 by Jyri Engeström and Petteri Koponen from Finland and launched in July of that year[7]. It was conceived based on the standard Short Message Service (SMS) becoming increasingly popular for communicating with family and close friends by its founders. Seeing its potential, Jaiku was later purchased by Google on October 9, 2007[8]. Engeström, one of its founders, previously discussed the concept of Social Objects applied to social media [Eng05].

**TWITTER** is an online news and social networking service on which users post and interact with messages known as *tweets*. Registered users can create tweets, but those who are unregistered can only read them. Users can access Twitter through its website interface, Short Message Service (SMS) or mobile application software. The service was created in March 2006 by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams and launched in July of that year [Bil14]. Similar to Jaiku, the idea of individuals using an SMS service to communicate with a small group was also the inspiration for Twitter. A landmark event for Twitter occurred during the South by Southwest Interactive (SXSWi) 2007 conference. Two 60-inches plasma televisions were placed in the conference hallways exclusively displaying Twitter messages in real-time streaming fashion. During this time, the usage of Twitter

5 https://www.omnicoreagency.com/twitter-statistics/
6 https://www.pcmag.com/article2/0,2704,2186676,00.asp
7 https://money.cnn.com/magazines/business2/business2_archive/2006/11/01/8392020/
8 https://web.archive.org/web/20071225101821/http://jaiku.com/help/google

increased from 20,000 to 60,000 tweets per day[9], and greatly helped to establish Twitter as one of the most widely known microblogging services in the world. Since its creation, Twitter has been evolving considerably during the years to stay relevant to the public by implementing new features such as media embedding and geo-location.

**TUMBLR** is a microblogging and social networking website that allows users to post multimedia and other content to a *short-form* blog[10]. Tumblr users can also make their blogs private if desired and many of the website's features are accessed from a *dashboard* interface. Tumblr was founded by David Karp in 2007, and owned by Oath Inc. On May 20, 2013, it was announced that Yahoo and Tumblr had reached an agreement for Yahoo! Inc. to purchase Tumblr for USD$ 1.1 billion in cash[11], evidencing the importance of Tumblr for social media. As of January 2016, the website had 555 million monthly visitors[12] and as of June 1, 2018, Tumblr hosts over 417.1 million blogs[13].

**POWNCE** was a free social networking and microblogging site centred on sharing messages, files, events, and links with friends. It was created by Internet entrepreneurs Kevin Rose, Leah Culver, and Daniel Burka[14]. At the time, the service was more feature-rich than Twitter. Pownce was launched on June 27, 2007, and was opened to the public on January 22, 2008. However, on December 1, 2008, Pownce announced that it had been acquired by blogging company Six Apart, and that the service would soon shut down[15], due to lack of revenue, stagnant growth, and an inability to compete with Twitter were the causes. It was subsequently shut down on December 15, 2008.

**FRIENDFEED** was a real-time feed aggregator that consolidated updates from social media and social networking websites, social bookmarking websites, blogs and microblogging updates, as well as any type of RSS/Atom feeds[16]. It was possible to use this stream of information to create customized feeds to share, as well as to originate new posts and discussions (including comments) with friends. According to their website, FriendFeed aimed to make content on the Web more relevant and useful by using existing social networks as a tool for discovering interesting information. FriendFeed was created by Bret Taylor, Jim Norris, Paul Buchheit and Sanjeev Singh, all former Google employees involved in services such as Gmail and Google Maps. On August 10, 2009, Facebook agreed to acquire it

9 https://thenextweb.com/twitter/2011/07/15/5-years-ago-today-twitter-launched-to-the-public/
10 https://gadgetwise.blogs.nytimes.com/2009/03/13/tumblr-makes-blogging-blissfully-easy/
11 https://money.cnn.com/2013/05/19/technology/yahoo-tumblr/index.html
12 https://www.alexa.com/siteinfo/tumblr.com
13 https://www.tumblr.com/about
14 https://www.crunchbase.com/organization/pownce
15 https://venturebeat.com/2008/12/01/six-apart-acquires-and-shuts-down-pownce/
16 https://web.archive.org/web/20110705022541/http://friendfeed.com/about/help

for USD$15 million in cash and USD$32.5 million in Facebook stock[17]. Facebook kept the service active until its closure on April 9, 2015.

**PLURK** is a free social networking and microblogging service that allows users to send updates (known as *plurks*) through short messages or links[18]. It was launched on May 12, 2008[19]. A distinctive feature of Plurk with respect to Twitter is called *karma*, the ability for users to score other users for social status and extra benefits such as more emoticons. Originally, messages were 140 characters long, like in Twitter, but was later increased to 210 and as of December 28, 2016, further raised to 360. Updates are displayed on the homepage of the users using a timeline, which lists all the updates received in chronological order. In addition, messages are delivered to other users who have chosen to receive them, in a publisher-subscriber fashion. Users can respond to updates of other users from their timeline through the Plurk website, by private or instant messaging, or by text messaging via compatible third party applications. According to Alexa, as of November 15, 2017, 70.1% of the traffic to Plurk comes from Taiwan, making this service a strong competitor to Twitter in this country[20].

**IDENTI.CA** is a free and open source social networking and microblogging service using the Activity Streams protocol[21]. Activity Streams is an open format specification used to syndicate activities taken in social web applications and services, e.g. Facebook, Instagram and Twitter. The Activity Streams protocol supports many kinds of activities such as games and is actively advocated by Gnip, a large multi-source social data aggregation company that was later purchased in April 2014 by Twitter[22]. Identi.ca uses the pump.io engine[23], a general purpose Activity Streams engine that can be used as a federated social networking protocol. Pump.io is a follow-up to the GNU Social server software, previously known as StatusNet and Laconica[24], which focuses on interoperability between social media services. Identi.ca is similar to social networking sites like Facebook and Google+, allowing unlimited length status updates, rich text and images. Originally, identi.ca did not support hashtags, global search or user groups, and stopped accepting new registrations in 2013.

17 https://techcrunch.com/2009/08/10/the-cost-of-friendfeed-roughly-50-million-in-cash-and-stock/
18 https://www.techinasia.com/what-is-plurk
19 https://web.archive.org/web/20080719135327/http://blog.plurk.com/2008/05/20/das-leben-der-anderen-a-window-into-the-lives-of-others/
20 https://www.alexa.com/siteinfo/plurk.com
21 https://activitystrea.ms/
22 https://www.reuters.com/article/us-twitter-gnip/twitter-buys-social-data-provider-gnip-stock-soars-idUSBREA3E17D20140415
23 http://pump.io/
24 https://web.archive.org/web/20090831073527/http://status.net/2009/08/28/laconica-is-now-statusnet/

**YAMMER** is an enterprise social networking service used for private communication within organizations[25]. The service is *freemium*, offering basic functionality for free but more advanced features can be obtained after payment. Access to a Yammer *network* is determined by the Internet domain of users so that only individuals with approved email addresses, e.g. belonging to their companies, may join their respective networks. Yammer was founded by David O. Sacks (a former Paypal employee) and Adam Pisoni. The service began as an internal communication system for the genealogy website Geni.com, and was launched as an independent product in 2008, being dubbed as the *Twitter for companies*[26]. Microsoft later acquired Yammer in 2012 for USD$1.2 billion in cash[27].

**SINA WEIBO** is a major Chinese microblogging service. The term *weibo* is the Chinese word for *microblog*. Sina Weibo was launched by the Sina Corporation (a large Chinese technology company) on August 14, 2009, and is one of the most popular social media platforms in China[28]. Many popular non-Chinese microblogging services such as Twitter, Facebook and Plurk were censored in the country at the time, however the Sina Corporation has enjoyed a long-standing government approval and took the opportunity for starting their own service[29]. In Sina Weibo, users can send messages and multimedia content through the website or a mobile application. They also can upload pictures and videos to the public for global sharing. Initially, Sina Weibo invited a large number of stars and celebrities to join the service. More recently it also invites media workers, government departments, enterprises, and non-governmental organizations. Similar to Twitter, accounts are clearly distinguished between verified celebrities and ordinary users. As of 2018, Sina Weibo has over 411 million monthly active users, with large stocks, advertising sales, revenue and total earnings[30], surpassing USD$30 billion market valuation mark for the first time.

**TENCENT WEIBO** is a Chinese microblogging (weibo) website launched by Tencent in April 2010 and a direct competitor to Sina Weibo in China[31]. Tencent is the world's fifth-largest internet company by revenue. Similar to Sina Weibo and Twitter, users can broadcast messages including 140 Chinese characters at most through the web, SMS or mobile application. Tencent Weibo was primarily conceived to curb the market competition and as of 2012, the service has had 469 million registered users[32].

25 https://www.pcworld.com/article/260517/what_is_heck_is_yammer.html
26 https://techcrunch.com/2008/09/08/yammer-launches-at-tc50-twitter-for-companies/
27 https://www.forbes.com/sites/shelisrael/2012/06/25/its-official-microsoft-buys-yammer-for-1-2-billion-cash/
28 https://www.chinadaily.com.cn/china/2011-03/02/content_12099500.htm
29 https://content.time.com/time/specials/packages/article/0,28804,2066367_2066369_2066392,00.html
30 https://www.investors.com/news/technology/weibo-reports-first-quarter-earnings/
31 https://techcrunch.com/2011/12/22/tencent-vs-sina-the-fight-for-chinas-social-graph/
32 https://www.techinasia.com/netease-weibo-260-million-users-numbers

**TOUT** is an innovative social networking and microblogging service that enables its users to send and view fifteen-second videos, known as *touts*[33]. Because of the exclusive video-oriented nature of the service, it is particularly popular with television journalists. The technology for the service was created at the Stanford Research Institute (SRI) International by Michael Downing based on two patents owned by SRI. Tout gained prominence in June 2011 when basketball star player Shaquille O'Neal used the service to announce his retirement[34]. As of early 2012, Tout has had over 12 million visitors since its launch, and 75 million touts have been created and shared by users of Tout[35].

**APP.NET** was an ad-free online social networking and microblogging service which enabled its users to write messages of up to 256 characters, dubbed the *developer-friendly Twitter*[36]. App.net provided their own web interface to the service, called *Alpha*, which was used by some users, however they highly encouraged the use and development of third-party applications instead. In contrast to Twitter, App.net had no advertising and relied on paid subscriptions for users and developers. The name *App.net* was originally used for a platform for application developers to showcase their applications, however on July 2012, Mixed Media Labs (founded by Dalton Caldwell) announced that App.net would change its purpose to be an ad-free social platform instead. Caldwell began directly crowd funding the system, with a goal of $500,000 and 10,000 backers[37]. On May 2014, it was announced that subscription renewals had been poor and the development staff could no longer be sustained, therefore future operations would be on a maintenance-only basis using contractors[38]. App.net finally closed as a social network on March 2017 and the source code for the service was made available on Github.

**TWISTER** is an experimental fully distributed microblogging platform[39]. It borrows technologies from Bitcoin (blockchain) and BitTorrent (file exchange)[40] to mimic traditional microblogging functionality. Therefore, it is known as the *distributed and secure Twitter clone*. Because the system is completely decentralised, theoretically no one is able to close it down, as there is no single point of control. Twister is designed to prevent other users from knowing geo-locations, IP addresses, and who the user is following. As with other microblogging platforms, users can publish public messages, however direct messages and private mes-

33 https://www.box.com/s/v7jt8v2qkl6zynpmamlb
34 https://www.espn.com/boston/nba/news/story?id=6615886
35 https://techcrunch.com/2012/03/14/tout-touts-plans/
36 https://www.technologyreview.com/s/428524/a-social-network-free-of-ads/
37 http://daltoncaldwell.com/an-audacious-proposal
38 https://web.archive.org/web/20170123034715/http://blog.app.net/2014/05/06/app-net-state-of-the-union/
39 http://techpresident.com/news/wegov/24759/making-NSA-proof-social-networking-mainstream
40 http://www.notebookreview.com/feature/cryptography-apps-how-to-keep-your-personal-info-private/

sages to other users are protected from unsolicited access. Twister originated in December 2013 from Brazilian computer engineer and programmer Miguel Freitas [Fre16], inspired after learning about the massive spy programs of the National Security Agency (NSA) of the United States as revealed by the well-known NSA whistle-blower Edward Snowden.

**GAB** is a controversial microblogging service created as an alternative to Twitter[41], promoting itself as supporting free speech. Gab is very similar to Twitter, allowing its users to read and write messages of up to 300 characters, called *gabs*. In addition, the service also offers multimedia functionality. Gab was created in August 2016 by Andrew Torba, citing *the entirely left-leaning Big Social monopoly* as part of the inspiration for the new service[42]. Torba also claims that he created Gab after reading reports that Facebook employees suppress conservative articles. After six months of beta testing, Gab entered open registrations in May 2017. Because Gab does not aim to censor the content in the platform, it attracted a large number of users of hate-speech to use the platform[43]. Consequently, Gab has been described as a platform for white supremacists and the alternative-right.

**MASTODON** is a distributed and federated social network with microblogging features similar to Twitter launched in October 2016[44]. Mastodon is part of a larger, interconnected and decentralized network of independent servers known as the Fediverse. In Mastodon, each individual server is known as an *instance* and can administrate its own rules, account privileges, and message visibility with other instances. Users join individual instances rather than a global website or application and can post short messages called *toots* for others to see, according to the privacy settings of the user and the particular instance they are using[45]. The flagship Mastodon instance is called *Mastodon.social* and as of early April 2017 it had about 42,000 users. Instances are often based on communal interests such as games, movies or technology. The adoption of Mastodon has risen from 766,500 users as of August 1, 2017, to one million users on December 1, 2017[46]. Mastodon is different from Twitter by its philosophy towards independently operated and moderated small hubs in contrast to a large centralised system[47]. Eugen Rochko (programmer and founder of Mastodon) believes that small, close communities can police toxic behaviour more effectively than a single small safety team of a large company[48].

41 https://www.theguardian.com/media/2016/nov/17/gab-alt-right-social-media-twitter
42 https://www.buzzfeednews.com/article/alexkantrowitz/new-social-network-gab-growing-fast-free-speech
43 https://www.newsweek.com/nazis-free-speech-hate-crime-jews-social-media-gab-weev-668614
44 https://au.pcmag.com/social-networking/47343/feature/what-is-mastodon-and-will-it-kill-twitter
45 https://www.theverge.com/2017/4/7/15183128/mastodon-open-source-twitter-clone-how-to-use
46 https://lou.lt/@mastodonusercount/99099871022836220
47 https://medium.com/tootsuite/learning-from-twitters-mistakes-c272d67bba76
48 https://au.pcmag.com/social-networking/47343/feature/what-is-mastodon-and-will-it-kill-twitter

**MICRO.BLOG** is an experimental microblogging and social networking service created by Manton Reece, who also authored the book *Indie Microblogging*. Indie is the short form for *independent*. The book emphasizes on big social media services such as Facebook or Twitter fully owning and hosting all the content created by its users and how difficult is for these users to move across platforms. Reece advocates instead on more independent publishing of social media content and created a set of tools for this purpose. Micro.blog was then launched on April 2017, after a successful Kickstarter campaign that reached its funding target within one day[49] The service is the first multi-user social media platform to support the Webmention and Micropub protocols standardized by the World Wide Web Consortium[50]. Micro.blog is built using Jekyll, a static website generation tool, and users can post using hosted accounts. Then, users use RSS feeds to export their posts and syndicate them into the network from other websites they run. Users can also import their posts from Twitter and the defunct microblogging service App.net. DreamHost, one of the largest web hosting companies in the host, also backed the Kickstarter campaign of Manton Reece and announced that they intent to help customers create independent microblogs hosted at DreamHost and that will be compatible with the micro.blog system[51].

All of the above services have contributed to the development of microblogging social networks. Pioneering platforms such as Jaiku and Twitter started establishing this new form of communication and shortly after many others followed, e.g. Tumblr, Pownce and FriendFeed. One of the first major milestones for microblogging was the acquisition of the original Jaiku service by Google in 2007, representing the first time a large-scale company demonstrated interest on this type of social network. Other services such as FriendFeed and Tumblr also attracted big players to purchase them, i.e. Facebook and Yahoo respectively. As time progressed, many services appeared mimicking or enhancing the original functionality of Jaiku and Twitter, highlighting the successful model of these services. A summary timeline can be seen in Table A.1, where the key events for each service described in this section are presented.

Many services experimented with different approaches over the same fundamental concept of microblogging. For example, Plurk introduced social status ranking, App.net pursued more technical users (developers), identi.ca adopted the Activity Streams protocol standard, and Twister focused on stronger security and privacy. Some services also introduced microblogging into

49 http://www.siliconhillsnews.com/2017/01/04/indie-microblogging-kickstarter-project-in-austin-reaches-its-goal-in-one-day/
50 https://indieweb.org/indieweb_network
51 https://www.dreamhost.com/blog/pitching-support-open-web/

**Table A.1:** Timeline of the Recent History of Microblogging Services.

| Year/Month | Event |
|---|---|
| 2005/10 | The term *tumblelog* is defined |
| 2006/02 | Jaiku is founded |
| 2006/07 | Twitter is launched |
| 2006/07 | Jaiku is launched |
| 2007/02 | Tumblr is launched |
| 2007/03 | Twitter is used during the SXSWi conference |
| 2007/06 | Pownce is launched |
| 2007/10 | Jaiku is purchased by Google |
| 2007/10 | Friendfeed is launched |
| 2008/05 | Plurk is launched |
| 2008/07 | identi.ca is launched based on Activity Streams |
| 2008/09 | Yammer is launched |
| 2008/12 | Pownce is merged into SixApart and closed |
| 2009/08 | FriendFeed is purchased by Facebook |
| 2009/08 | Sina Weibo is launched |
| 2010/04 | Tencent Weibo is launched |
| 2010/04 | Tout is launched |
| 2012/01 | Jaiku is closed |
| 2012/08 | App.net is launched |
| 2012/12 | Twitter reports over 35 million users in China |
| 2013/05 | Tumblr is purchased by Yahoo |
| 2013/07 | identi.ca stopped accepting new registrations |
| 2013/12 | Twister is released |
| 2014/04 | Activity Streams adopted by Twitter |
| 2015/04 | FriendFeed is closed |
| 2016/08 | Gab is launched for beta testing |
| 2016/10 | Mastodon is released |
| 2016/12 | Plurk increased size limit to 360 characters |
| 2017/03 | App.net is closed |
| 2017/04 | micro.blog is launched |
| 2017/05 | Gab is launched for open registration |
| 2017/11 | Twitter increased size limit to 280 characters |

Table A.2: Selection of Microblogging Services from Section A.1 and how they relate to Twitter.

| Service | Comparison to Twitter |
|---|---|
| Tumblr | Posts focused on multimedia, e.g. photos, videos or audio |
| Plurk | Social scoring for its users, i.e. *karma* |
| Yammer | Microblogging workflow designed for companies |
| Sina Weibo | Twitter clone for the Chinese market |
| App.net | Framework for developers supported with subscriptions |
| Twister | Decentralised and more secure architecture |
| Gab | Uncensored Twitter clone, i.e. *free speech* |
| Mastodon | Community governed philosophy instead of centralised management |
| Micro.blog | Advocate users to own their content instead of the platform |

their home countries, e.g. Sina Weibo and Tencent in China (notorious for the reclusive regimen regarding online access for its citizen) and Plurk in Taiwan.

## A.2 TWITTER AS MICROBLOGGING REFERENCE

Twitter has remained as one of the most successful microblogging platforms in the world since its breakthrough during the SXSWi in 2007. However, a number of the services described in Section A.1 can be also identified to be specially related to Twitter as seen in Table A.2.

Despite the effort of some services to capitalise over the success of Twitter, not all of them could survive the rapidly changing landscape of microblogging networks. Even though these alternative platforms proposed interesting enhancements to the baseline established by Twitter, some services were absorbed or simply disappeared. For instance, Pownce was more feature-rich than Twitter in 2007, however it closed due to lack of revenue for the company. Similarly, App.net was oriented to a growing developer community but its subscription renewals were poor. A demonstration of the increasing influence of microblogging services is the case of Gab, which proposed a free speech version of Twitter. Unfortunately, the uncensored nature of Gab also attracted controversial participants such as white supremacists and hate speech.

Microblogging services are proving to be important players in modern social media. Users seem to highly value the ability to quickly broadcast opinions, thoughts and daily happenings through easy to use and expressive mechanisms such as short text, images, audio or links. Furthermore, a significant amount of research has also been carried in the context of microblogging as is described in Chapter 2. Overall, Twitter has become the reference point for microblogging services and therefore this thesis adopts it as the representative service for experimentation.

# B STRUCTURAL PROPERTIES

In this appendix, the structural properties of the ground-truth functional communities defined in Chapter 3 are presented for both, the static and the dynamic scenario.

Each structural property $p$ in the tables below is computed for every ground-truth community $C_i$ and its corresponding non-community $\tilde{C}_i$, and then the average ratio $r = p(C_i)/p(\tilde{C}_i)$ is reported in the tables for all community types in each experimental dataset. If $r > 1.0$, then a measurable difference in the structural property $p$ for $C_i$ compared to $\tilde{C}_i$ can be asserted.

The structural properties are: (1) CC, Clustering Coefficient, (2) AvgDeg, Average Degree, (3) Density, Edge Density, and (4) Cohesiv, Cohesiveness.

## B.1 STRUCTURAL PROPERTIES IN THE STATIC SCENARIO

**Table B.1:** Ratio between structural properties of ground-truth functional communities and randomly chosen nodes with similar shortest path distribution for the Pope2013 dataset.

| C. Type | CC | AvgDeg | Density | Cohesiv | All $> 1.0$ |
|---|---|---|---|---|---|
| cities | 1.0000 | 1.0000 | 1.0000 | 1.0000 | |
| countries | 0.2299 | 0.9110 | 0.8604 | 0.3269 | |
| hashtags | 0.7806 | 0.9941 | 0.9838 | 0.7996 | |
| mentions | 2.5029 | 1.1644 | 1.1265 | 2.3654 | Yes |
| places | $\infty$ | 1.1429 | 1.1667 | 1.7500 | Yes |
| retweets | 0.6327 | 0.9810 | 0.9713 | 0.6847 | |
| urls | 0.7920 | 0.9946 | 0.9853 | 0.7954 | |
| **Average** | **0.9897** | **1.0269** | **1.0134** | **1.1032** | |

Table B.2: Ratio between structural properties of ground-truth functional communities and randomly chosen nodes with similar shortest path distribution for the POPE2013-SPL dataset.

| C. Type | CC | AvgDeg | Density | Cohesiv | All > 1.0 |
|---|---|---|---|---|---|
| cities | 0.0000 | 0.9888 | 0.9879 | 0.0000 | |
| countries | 2.7000 | 1.0090 | 1.0125 | 2.2500 | Yes |
| hashtags | 1.9335 | 1.0158 | 1.0102 | 1.9450 | Yes |
| mentions | 11.0947 | 1.1539 | 1.1363 | 9.9305 | Yes |
| places | 0.0000 | 0.9826 | 0.9867 | 0.0000 | |
| retweets | 4.0359 | 1.0928 | 1.0535 | 4.1643 | Yes |
| urls | 1.9030 | 1.0086 | 1.0103 | 1.8845 | Yes |
| **Average** | **3.0953** | **1.0359** | **1.0282** | **2.8820** | **Yes** |

Table B.3: Ratio between structural properties of ground-truth functional communities and randomly chosen nodes with similar shortest path distribution for the WORLDCUP2014 dataset.

| C. Type | CC | AvgDeg | Density | Cohesiv | All > 1.0 |
|---|---|---|---|---|---|
| cities | 0.6344 | 0.9725 | 0.9687 | 0.7159 | |
| countries | 0.4352 | 0.9627 | 0.9591 | 0.4275 | |
| hashtags | 1.0791 | 1.0338 | 1.0012 | 1.1011 | Yes |
| mentions | 2.8791 | 1.3235 | 1.2046 | 2.4441 | Yes |
| places | 0.7645 | 0.9871 | 0.9773 | 0.7187 | |
| retweets | 1.2342 | 1.0637 | 1.0121 | 1.2003 | Yes |
| urls | 1.4062 | 1.0579 | 1.0193 | 1.3630 | Yes |
| **Average** | **1.2047** | **1.0573** | **1.0203** | **1.1387** | **Yes** |

Table B.4: Ratio between structural properties of ground-truth functional communities and randomly chosen nodes with similar shortest path distribution for the RTE2015 dataset.

| C. Type | CC | AvgDeg | Density | Cohesiv | All > 1.0 |
|---|---|---|---|---|---|
| cities | 0.4034 | 1.0178 | 0.9169 | 0.5489 | |
| countries | 0.4365 | 0.9958 | 0.9510 | 0.6912 | |
| hashtags | 2.1117 | 1.2542 | 1.0885 | 2.0287 | Yes |
| mentions | 3.7619 | 1.7942 | 1.3538 | 3.1683 | Yes |
| places | 0.3981 | 0.9914 | 0.9329 | 0.4795 | |
| quotes | 2.3291 | 1.3907 | 1.1491 | 2.2839 | Yes |
| retweets | 2.8460 | 1.6003 | 1.1834 | 2.6283 | Yes |
| urls | 2.6746 | 1.2983 | 1.1495 | 2.4909 | Yes |
| **Average** | **1.8702** | **1.2928** | **1.0906** | **1.7900** | **Yes** |

Table B.5: Ratio between structural properties of ground-truth functional communities and randomly chosen nodes with similar shortest path distribution for the IRELAND2017 dataset.

| C. Type | CC | AvgDeg | Density | Cohesiv | All $> 1.0$ |
|---|---|---|---|---|---|
| cities | 22.2567 | 1.1828 | 1.0691 | 12.4548 | Yes |
| countries | 8.6811 | 1.0263 | 1.0205 | 6.8682 | Yes |
| hashtags | 31.8735 | 1.2197 | 1.1139 | 15.7759 | Yes |
| mentions | 48.7442 | 1.4785 | 1.2394 | 22.7188 | Yes |
| places | 11.1738 | 1.0794 | 1.0380 | 7.5388 | Yes |
| quotes | 38.2470 | 1.3039 | 1.1757 | 20.3443 | Yes |
| **Average** | **26.8294** | **1.2151** | **1.1094** | **14.2835** | **Yes** |

## B.2 STRUCTURAL PROPERTIES IN THE DYNAMIC SCENARIO

Table B.6: Ratio between structural properties of ground-truth functional communities and randomly chosen nodes with similar shortest path distribution for the POPE2013 dataset.

| C. Type | CC | AvgDeg | Density | Cohesiv | All $> 1.0$ |
|---|---|---|---|---|---|
| cities | $\infty$ | 1.2121 | 1.2000 | $\infty$ | Yes |
| countries | $\infty$ | 1.0354 | 1.0283 | $\infty$ | Yes |
| hashtags | 0.8872 | 1.0004 | 0.9905 | 0.9102 | |
| mentions | 2.4437 | 1.1659 | 1.1293 | 2.2365 | Yes |
| places | 2.7143 | 1.1111 | 1.1250 | 2.3333 | Yes |
| retweets | 0.8632 | 1.0073 | 0.9845 | 0.9850 | |
| urls | 0.8626 | 0.9992 | 0.9900 | 0.8737 | |
| **Average** | **1.5542** | **1.0759** | **1.0639** | **1.4677** | **Yes** |

Table B.7: Ratio between structural properties of ground-truth functional communities and randomly chosen nodes with similar shortest path distribution for the POPE2013-SPL dataset.

| C. Type | CC | AvgDeg | Density | Cohesiv | All $> 1.0$ |
|---|---|---|---|---|---|
| cities | $\infty$ | 1.0000 | 1.0000 | $\infty$ | |
| countries | 2.3410 | 1.0067 | 1.0115 | 1.7697 | Yes |
| hashtags | 2.2368 | 1.0224 | 1.0147 | 2.2821 | Yes |
| mentions | 12.0279 | 1.1830 | 1.1590 | 10.5116 | Yes |
| places | 0.0000 | 0.9826 | 0.9811 | 0.0000 | |
| retweets | 5.9681 | 1.1056 | 1.0667 | 5.5784 | Yes |
| urls | 1.8435 | 1.0094 | 1.0096 | 1.9421 | Yes |
| **Average** | **4.0696** | **1.0443** | **1.0347** | **3.6806** | **Yes** |

**Table B.8:** Ratio between structural properties of ground-truth functional communities and randomly chosen nodes with similar shortest path distribution for the WorldCup2014 dataset.

| C. Type | CC | AvgDeg | Density | Cohesiv | All $> 1.0$ |
|---|---|---|---|---|---|
| cities | 1.1278 | 1.0036 | 0.9934 | 1.1367 | |
| countries | 0.6844 | 0.9731 | 0.9707 | 0.6618 | |
| hashtags | 1.6123 | 1.1167 | 1.0407 | 1.6710 | Yes |
| mentions | 3.2094 | 1.3688 | 1.2474 | 2.7421 | Yes |
| places | 0.9951 | 0.9872 | 0.9870 | 0.9805 | |
| retweets | 1.8448 | 1.1405 | 1.0577 | 1.7257 | Yes |
| urls | 1.8492 | 1.1047 | 1.0506 | 1.7278 | Yes |
| **Average** | **1.6176** | **1.0992** | **1.0496** | **1.5208** | **Yes** |

**Table B.9:** Ratio between structural properties of ground-truth functional communities and randomly chosen nodes with similar shortest path distribution for the RTE2015 dataset.

| C. Type | CC | AvgDeg | Density | Cohesiv | All $> 1.0$ |
|---|---|---|---|---|---|
| cities | 5.2716 | 1.2251 | 1.1334 | 4.9728 | Yes |
| countries | 2.4316 | 1.2016 | 1.1400 | 2.4288 | Yes |
| hashtags | 3.5531 | 1.8086 | 1.3510 | 3.0795 | Yes |
| mentions | 4.2429 | 2.2520 | 1.5515 | 3.4606 | Yes |
| places | 3.0567 | 1.1822 | 1.0989 | 3.2423 | Yes |
| quotes | 3.8770 | 1.9087 | 1.5038 | 3.2801 | Yes |
| retweets | 4.0560 | 2.1774 | 1.5220 | 3.3237 | Yes |
| urls | 3.3876 | 1.5076 | 1.2977 | 3.0178 | Yes |
| **Average** | **3.7345** | **1.6579** | **1.3248** | **3.3507** | **Yes** |

**Table B.10:** Ratio between structural properties of ground-truth functional communities and randomly chosen nodes with similar shortest path distribution for the Ireland2017 dataset.

| C. Type | CC | AvgDeg | Density | Cohesiv | All $> 1.0$ |
|---|---|---|---|---|---|
| cities | 79.8334 | 2.1928 | 1.2812 | 18.3795 | Yes |
| countries | 32.1232 | 2.2361 | 1.0373 | 7.8326 | Yes |
| hashtags | 63.4964 | 1.6238 | 1.3834 | 30.7773 | Yes |
| mentions | 78.9327 | 1.9163 | 1.5747 | 37.6096 | Yes |
| places | 79.6753 | 2.0145 | 1.1689 | 18.2734 | Yes |
| quotes | 68.8420 | 1.5843 | 1.4132 | 34.8451 | Yes |
| **Average** | **67.1505** | **1.9280** | **1.3098** | **24.6196** | **Yes** |

# BIBLIOGRAPHY


[AEC+15] M. Ayati, S. Erten, M. R. Chance, and M. Koyutürk. "MOBAS: identification of disease-associated protein subnetworks using modularity-based scoring". In: *EURASIP Journal on Bioinformatics and Systems Biology* (2015). ISSN: 1687-4145.

[AH17] Y. Amichai-Hamburger and T. Hayat. "Social Networking". In: *The International Encyclopedia of Media Effects*. American Cancer Society, 2017, pp. 1–12. ISBN: 978-1-118-78376-4.

[ARL18] U. Aslak, M. Rosvall, and S. Lehmann. "Constrained information flows in temporal networks reveal intermittent communities". In: *Physical Review E* (2018), p. 062312.

[Asl18] S. Aslam. *Twitter by the Numbers (2018): Stats, Demographics & Fun Facts*. 2018.

[AVC+16] B. Amor, S. Vuik, R. Callahan, A. Darzi, S. N. Yaliraki, and M. Barahona. *Community detection and role identification in directed networks: understanding the Twitter network of the care.data debate*. World Scientific, 2016. ISBN: 978-1-60558-752-3.

[AW10] *Managing and Mining Graph Data*. Springer US, 2010. ISBN: 978-1-4419-6044-3.

[BA99] A.-L. Barabási and R. Albert. "Emergence of Scaling in Random Networks". In: *Science* (1999), pp. 509–512. ISSN: 0036-8075, 1095-9203.

[Bar03] A.-L. Barabási. "Linked: The New Science of Networks". In: *American Journal of Physics* (2003), pp. 409–410. ISSN: 0002-9505.

[Bar54] J. A. Barnes. "Class and Committees in a Norwegian Island Parish". In: *Human Relations* (1954), pp. 39–58. ISSN: 0018-7267.

[BBB+08] L. Barkhuus, B. Brown, M. Bell, S. Sherwood, M. Hall, and M. Chalmers. "From Awareness to Repartee: Sharing Location Within Social Groups". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2008, pp. 497–506. ISBN: 978-1-60558-011-1.

[BGL+08] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* (2008), P10008. ISSN: 1742-5468.

[BGL10] D. Boyd, S. Golder, and G. Lotan. "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter". In: *2010 43rd Hawaii International Conference on System Sciences*. 2010 43rd Hawaii International Conference on System Sciences. 2010, pp. 1–10.

[BHH+16] A. Barraza-Urbina, H. Hromic, B. Heitmann, A. Yañez, H. Tamatam, and C. Hayes. *Using Social Media for Online Television Adaptation Services at RTÉ Ireland*. Technical Report. Insight Centre for Data Analytics, National University of Ireland, Galway, 2016.

[Bil14] N. Bilton. *Hatching Twitter: A true story of money, power, friendship, and betrayal*. Penguin, 2014.

[BLH12] V. Belák, S. Lam, and C. Hayes. "Cross-Community Influence in Discussion Fora". In: *Sixth International AAAI Conference on Weblogs and Social Media*. Sixth International AAAI Conference on Weblogs and Social Media. 2012.

[BLL15] M. Bakillah, R.-Y. Li, and S. H. L. Liang. "Geo-located community detection in Twitter with enhanced fast-greedy optimization of modularity: the case study of typhoon Haiyan". In: *International Journal of Geographical Information Science* (2015), pp. 258–279. ISSN: 1365-8816.

[BMB+09] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca. "Network Analysis in the Social Sciences". In: *Science* (2009), pp. 892–895. ISSN: 0036-8075, 1095-9203.

[Bon76] J. A. Bondy. *Graph Theory With Applications*. Elsevier Science Ltd., 1976. ISBN: 978-0-444-19451-0.

[Bra02] D. Braha. "Partitioning Tasks to Product Development Teams". In: (2002), pp. 333–344.

[CD11] S. J. Cranmer and B. A. Desmarais. "Inferential Network Analysis with Exponential Random Graph Models". In: *Political Analysis* (2011), pp. 66–86. ISSN: 1047-1987, 1476-4989.

[CG97] F. R. K. Chung and F. C. Graham. *Spectral Graph Theory*. American Mathematical Soc., 1997. 228 pp. ISBN: 978-0-8218-0315-8.

[CGP11] M. Coscia, F. Giannotti, and D. Pedreschi. "A classification for community discovery methods in complex networks". In: *Statistical Analysis and Data Mining* (2011), pp. 512–546. ISSN: 1932-1872.

[CHD10] J. Chan, C. Hayes, and E. M. Daly. "Decomposing Discussion Forums and Boards Using User Roles". In: *Fourth International AAAI Conference on Weblogs and Social Media*. Fourth International AAAI Conference on Weblogs and Social Media. 2010.

[CK01] A. Condon and R. M. Karp. "Algorithms for graph partitioning on the planted partition model". In: *Random Structures & Algorithms* (2001), pp. 116–140. ISSN: 1098-2418.

[CMG09] M. Cha, A. Mislove, and K. P. Gummadi. "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, pp. 721–730. ISBN: 978-1-60558-487-4.

[CNM04] A. Clauset, M. E. J. Newman, and C. Moore. "Finding community structure in very large networks". In: *Physical Review E* (2004), p. 066111.

[CRF+11] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Goncalves, F. Menczer, and A. Flammini. "Political Polarization on Twitter". In: *Fifth International AAAI Conference on Weblogs and Social Media*. Fifth International AAAI Conference on Weblogs and Social Media. 2011.

[CT12] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012. 688 pp. ISBN: 978-1-118-58577-1.

[DFC05] J. Diesner, T. L. Frantz, and K. M. Carley. "Communication Networks from the Enron Email Corpus "It's Always About the People. Enron is no Different"". In: *Computational & Mathematical Organization Theory* (2005), pp. 201–228. ISSN: 1572-9346.

[DOG15] D. Darmon, E. Omodei, and J. Garland. "Followers Are Not Enough: A Multifaceted Approach to Community Detection in Online Social Networks". In: *PLOS ONE* (2015), e0134860. ISSN: 1932-6203.

[DSJ+10] M. De Choudhury, H. Sundaram, A. John, D. D. Seligmann, and A. Kelliher. *"Birds of a Feather": Does User Homophily Impact Information Diffusion in Social Media?* 2010.

[EK10] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010. 745 pp. ISBN: 978-1-139-49030-6.

[Eng05] J. Engeström. *Why some social network services work and others don't—Or: the case for object-centered sociality*. Zengestrom. 2005.

[ER59] P. Erdős and A. Rényi. "On random graphs". In: *Publicationes Mathematicae Debrecen* (1959), pp. 290–297.

[Eva96] J. D. Evans. *Straightforward Statistics for the Behavioral Sciences*. Brooks/Cole Publishing Company, 1996. 632 pp. ISBN: 978-0-534-23100-2.

[FB07] S. Fortunato and M. Barthélemy. "Resolution limit in community detection". In: *Proceedings of the National Academy of Sciences* (2007), pp. 36–41. ISSN: 0027-8424, 1091-6490.

[FB17] H. Fani and E. Bagheri. "Community detection in social networks". In: *Encyclopedia with Semantic Computing and Robotic Intelligence* (2017), p. 1630001. ISSN: 2529-7376.

[FH16] S. Fortunato and D. Hric. "Community detection in networks: A user guide". In: *Physics Reports* (2016), pp. 1–44. ISSN: 0370-1573.

[FLG00] G. W. Flake, S. Lawrence, and C. L. Giles. "Efficient Identification of Web Communities". In: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2000, pp. 150–160. ISBN: 978-1-58113-233-5.

[For10] S. Fortunato. "Community detection in graphs". In: *Physics Reports* (2010), pp. 75–174. ISSN: 0370-1573.

[Fre04] L. Freeman. "The development of social network analysis". In: *A Study in the Sociology of Science* (2004).

[Fre16] M. Freitas. "Twister: The Development of a Peer-to-peer Microblogging Platform". In: *Int. J. Parallel Emerg. Distrib. Syst.* (2016), pp. 20–33. ISSN: 1744-5760.

[GA05] R. Guimerà and L. A. N. Amaral. "Cartography of complex networks: modules and universal roles". In: *Journal of Statistical Mechanics: Theory and Experiment* (2005), P02001. ISSN: 1742-5468.

[GDC10] D. Greene, D. Doyle, and P. Cunningham. "Tracking the Evolution of Communities in Dynamic Social Networks". In: *2010 International Conference on Advances in Social Networks Analysis and Mining*. 2010 International Conference on Advances in Social Networks Analysis and Mining. 2010, pp. 176–183.

[GHW97] L. Garton, C. Haythornthwaite, and B. Wellman. "Studying Online Social Networks". In: *Journal of Computer-Mediated Communication* (1997).

[GJK12] A. Gupta, A. Joshi, and P. Kumaraguru. "Identifying and Characterizing User Communities on Twitter During Crisis Events". In: *Proceedings of the 2012 Workshop on DUBMMSM*. ACM, 2012, pp. 23–26. ISBN: 978-1-4503-1707-8.

[GN02] M. Girvan and M. E. J. Newman. "Community structure in social and biological networks". In: *Proceedings of the National Academy of Sciences* (2002), pp. 7821–7826. ISSN: 0027-8424, 1091-6490.

[Gra76] M. Granovetter. "Network Sampling: Some First Steps". In: *American Journal of Sociology* (1976), pp. 1287–1303. ISSN: 0002-9602.

[GWT11] A. Gruzd, B. Wellman, and Y. Takhteyev. "Imagining Twitter as an Imagined Community". In: *American Behavioral Scientist* (2011), pp. 1294–1318. ISSN: 0002-7642.

[HBG+14] S. Harenberg, G. Bello, L. Gjeltema, S. Ranshous, J. Harlalka, R. Seay, K. Padmanabhan, and N. Samatova. "Community detection in large-scale networks: a survey and empirical evaluation". In: *Wiley Interdisciplinary Reviews: Computational Statistics* (2014), pp. 426–439. ISSN: 1939-0068.

[HBH+17] H. Hromic, A. Barraza-Urbina, C. Hayes, and N. Cantle. "Mining TV Twitter Networks for Adaptive Content Navigation and Community Awareness". In: *Expert Update* (2017).

[HH14] H. Hromic and C. Hayes. "Constructing Twitter Datasets using Signals for Event Detection Evaluation". In: *Synergies of Case-Based Reasoning and Data Mining Workshop*. 22nd International Conference on Case-Based Reasoning. 2014.

[HH18a] H. Hromic and C. Hayes. "Characterising and Evaluating Online Communities from Live Microblogging User Interactions". In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2018, pp. 21–24.

[HH18b] H. Hromic and C. Hayes. "Visualising the Evolution of Dynamic Communities in Social Networks using Timelines". In: *3nd ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data (AALTD)*. The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD). 2018.

[HHS+11] B. Hecht, L. Hong, B. Suh, and E. H. Chi. "Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 237–246. ISBN: 978-1-4503-0228-9.

[HKW+12] H. Hromic, M. Karnstedt, M. Wang, A. Hogan, V. Belák, and C. Hayes. "Event Panning in a Stream of Big Data". In: *LWA Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML)*. 2012.

[HPH+15] H. Hromic, N. Prangnawarat, I. Hulpuş, M. Karnstedt, and C. Hayes. "Graph-Based Methods for Clustering Topics of Interest in Twitter". In: *Engineering the Web in the Big Data Era*. Springer International Publishing, 2015, pp. 701–704. ISBN: 978-3-319-19890-3.

[HSS00] P. Hedström, R. Sandell, and C. Stern. "Mesolevel Networks and the Diffusion of Social Movements: The Case of the Swedish Social Democratic Party". In: *American Journal of Sociology* (2000), pp. 145–172. ISSN: 0002-9602.

[IH93] B. Iglewicz and D. C. Hoaglin. *How to Detect and Handle Outliers*. ASQC Quality Press, 1993. 108 pp. ISBN: 978-0-87389-247-6.

[Jam06] P. James. *Globalism, Nationalism, Tribalism: Bringing Theory Back in*. Pine Forge Press, 2006. 384 pp. ISBN: 978-1-4462-3054-1.

[Jav08] A. Java. "Mining Social Media Communities and Content". PhD thesis. University of Maryland at Baltimore County, 2008.

[JNH+12] P. James, Y. Nadarajah, K. Haive, and V. Stead. *Sustainable Communities, Sustainable Development: Other Paths for Papua New Guinea*. University of Hawai'i Press, 2012. ISBN: 978-0-8248-6120-9.

[JOP+01] E. Jones, T. Oliphant, P. Peterson, et al. *SciPy: Open source scientific tools for Python*. 2001.

[JSF+07] A. Java, X. Song, T. Finin, and B. Tseng. "Why We Twitter: Understanding Microblogging Usage and Communities". In: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*. ACM, 2007, pp. 56–65. ISBN: 978-1-59593-848-0.

[JSF+09] A. Java, X. Song, T. Finin, and B. Tseng. "Why We Twitter: An Analysis of a Microblogging Community". In: *Advances in Web Mining and Web Usage Analysis*. Springer Berlin Heidelberg, 2009, pp. 118–138. ISBN: 978-3-642-00528-2.

[JV11] C. Jones and E. H. Volpe. "Organizational identification: Extending our understanding of social identities through social networks". In: *Journal of Organizational Behavior* (2011), pp. 413–434. ISSN: 1099-1379.

[Kad12] C. Kadushin. *Understanding Social Networks: Theories, Concepts, and Findings*. Oxford University Press, USA, 2012. 266 pp. ISBN: 978-0-19-537947-1.

[KCH+13] A. Kan, J. Chan, C. Hayes, B. Hogan, J. Bailey, and C. Leckie. "A time decoupling approach for studying forum dynamics". In: *World Wide Web* (2013), pp. 595–620. ISSN: 1573-1413.

[KHM+11] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre. "Social media? Get serious! Understanding the functional building blocks of social media". In: *Business Horizons* (2011), pp. 241–251. ISSN: 0007-6813.

[KLP+10] H. Kwak, C. Lee, H. Park, and S. Moon. "What is Twitter, a Social Network or a News Media?" In: *Proceedings of the 19th International Conference on World Wide Web*. ACM, 2010, pp. 591–600. ISBN: 978-1-60558-799-8.

[KNT10] R. Kumar, J. Novak, and A. Tomkins. "Structure and Evolution of Online Social Networks". In: *Link Mining: Models, Algorithms, and Applications*. Springer New York, 2010, pp. 337–357. ISBN: 978-1-4419-6515-8.

[KRC+11] M. Karnstedt, M. Rowe, J. Chan, H. Alani, and C. Hayes. "The Effect of User Features on Churn in Social Networks". In: *Proceedings of the 3rd International Web Science Conference*. ACM, 2011, 23:1–23:8. ISBN: 978-1-4503-0855-7.

[LB14] X. Lu and C. Brelsford. "Network Structure and Community Evolution on Twitter: Human Behavior Change in Response to the 2011 Japanese Earthquake and Tsunami". In: *Scientific Reports* (2014), p. 6773. ISSN: 2045-2322.

[LC14] C. Lee and P. Cunningham. "Community detection: effective evaluation on large social networks". In: *Journal of Complex Networks* (2014), pp. 19–37. ISSN: 2051-1310.

[LFR08] A. Lancichinetti, S. Fortunato, and F. Radicchi. "Benchmark graphs for testing community detection algorithms". In: *Physical Review E* (2008), p. 046110.

[LLD+08] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. "Statistical Properties of Community Structure in Large Social and Information Networks". In: *Proceedings of the 17th International Conference on World Wide Web*. ACM, 2008, pp. 695–704. ISBN: 978-1-60558-085-2.

[LLM10] J. Leskovec, K. J. Lang, and M. Mahoney. "Empirical comparison of algorithms for network community detection". In: *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 631–640.

[LM54] P. F. Lazarsfeld and R. K. Merton. "Friendship as a social process: A substantive and methodological analysis". In: *Freedom and control in modern society* (1954), pp. 18–66.

[LRR+11] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato. "Finding Statistically Significant Communities in Networks". In: *PLoS ONE* (2011). ISSN: 1932-6203.

[Lux07] U. v. Luxburg. "A tutorial on spectral clustering". In: *Statistics and Computing* (2007), pp. 395–416. ISSN: 0960-3174, 1573-1375.

[Mar88] P. V. Marsden. "Homogeneity in confiding relations". In: *Social Networks* (1988), pp. 57–76. ISSN: 0378-8733.

[MC86] D. W. McMillan and D. M. Chavis. "Sense of community: A definition and theory". In: *Journal of Community Psychology* (1986), pp. 6–23. ISSN: 1520-6629.

[MDC+16] F. Musciotto, S. Delpriori, P. Castagno, and E. Pournaras. "Mining Social Interactions in Privacy-preserving Temporal Networks". In: *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press, 2016, pp. 1103–1110. ISBN: 978-1-5090-2846-7.

[ML14] J. Mcauley and J. Leskovec. "Discovering Social Circles in Ego Networks". In: *ACM Trans. Knowl. Discov. Data* (2014), 4:1–4:28. ISSN: 1556-4681.

[MPC+06] A. A. Moreira, D. R. Paula, R. N. Costa Filho, and J. S. Andrade. "Competitive cluster growth in complex networks". In: *Physical Review E* (2006), p. 065101.

[MPL+13] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley. "Is the sample good enough? Comparing data from twitter's streaming API with Twitter's firehose". In: 7th International AAAI Conference on Weblogs and Social Media, ICWSM 2013. AAAI press, 2013.

[MSC01] M. McPherson, L. Smith-Lovin, and J. M. Cook. "Birds of a Feather: Homophily in Social Networks". In: *Annual Review of Sociology* (2001), pp. 415–444. ISSN: 0360-0572.

[MV13] F. D. Malliaros and M. Vazirgiannis. "Clustering and community detection in directed networks: A survey". In: *Physics Reports* (2013), pp. 95–142. ISSN: 0370-1573.

[NBW11] M. Newman, A.-L. Barabási, and D. J. Watts. *The Structure and Dynamics of Networks*. Princeton University Press, 2011. 593 pp. ISBN: 978-1-4008-4135-6.

[New03] M. Newman. "The Structure and Function of Complex Networks". In: *SIAM Review* (2003), pp. 167–256. ISSN: 0036-1445.

[New06a] M. E. Newman. "Modularity and community structure in networks". In: *Proceedings of the National Academy of Sciences of the United States of America* (2006), pp. 8577, 8577–8582. ISSN: 0027-8424.

[New06b] M. E. J. Newman. "Finding community structure in networks using the eigenvectors of matrices". In: *Physical Review E* (2006), p. 036104.

[NG04] M. E. J. Newman and M. Girvan. "Finding and evaluating community structure in networks". In: *Physical Review E* (2004), p. 026113.

[Noo14] W. de Nooy. "Social Network Analysis, Graph Theoretical Approaches to". In: *Encyclopedia of Complexity and Systems Science*. Springer Berlin Heidelberg, 2014, pp. 1–20. ISBN: 978-3-642-27737-5.

[NP03] M. E. J. Newman and J. Park. "Why social networks are different from other types of networks". In: *Physical Review E* (2003), p. 036122.

[Par18] K. I. Park. *Fundamentals of Probability and Stochastic Processes with Applications to Communications*. Springer International Publishing, 2018. ISBN: 978-3-319-68074-3.

[PBM+99] L. Page, S. Brin, R. Motwani, and T. Winograd. *The PageRank Citation Ranking: Bringing Order to the Web.* Techreport. 1999.

[PBV07] G. Palla, A.-L. Barabási, and T. Vicsek. "Quantifying social group evolution". In: *Nature* (2007), pp. 664–667. ISSN: 1476-4687.

[PDF+05] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. "Uncovering the overlapping community structure of complex networks in nature and society". In: *Nature* (2005), pp. 814–818. ISSN: 1476-4687.

[PKV+11] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos. "Community detection in Social Media". In: *Data Mining and Knowledge Discovery* (2011), pp. 515–554. ISSN: 1384-5810, 1573-756X.

[POM09] M. A. Porter, J.-P. Onnela, and P. J. Mucha. "Communities in Networks". In: *Notices of the AMS* (2009), pp. 1082–1097.

[PRS11] S. Parthasarathy, Y. Ruan, and V. Satuluri. "Community Discovery in Social Networks: Applications, Methods and Emerging Trends". In: *Social Network Data Analytics*. Springer US, 2011, pp. 79–113. ISBN: 978-1-4419-8461-6.

[PW10] O. Pele and M. Werman. "The Quadratic-Chi Histogram Distance Family". In: *Computer Vision – ECCV 2010*. European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2010, pp. 749–762. ISBN: 978-3-642-15551-2.

[PZ17] M. Q. Pasta and F. Zaidi. "Topology of Complex Networks and Performance Limitations of Community Detection Algorithms". In: *IEEE Access* (2017), pp. 10901–10914. ISSN: 2169-3536.

[RAK07] U. N. Raghavan, R. Albert, and S. Kumara. "Near linear time algorithm to detect community structures in large-scale networks". In: *Physical Review E* (2007), p. 036106.

[RB07] M. Rosvall and C. T. Bergstrom. "An information-theoretic framework for resolving community structure in complex networks". In: *Proceedings of the National Academy of Sciences* (2007), pp. 7327–7331. ISSN: 0027-8424, 1091-6490.

[RB08] M. Rosvall and C. T. Bergstrom. "Maps of random walks on complex networks reveal community structure". In: *Proceedings of the National Academy of Sciences* (2008), pp. 1118–1123. ISSN: 0027-8424, 1091-6490.

[RCC+04] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. "Defining and identifying communities in networks". In: *Proceedings of the National Academy of Sciences of the United States of America* (2004), pp. 2658–2663. ISSN: 0027-8424, 1091-6490.

[RHM08] W. Ryan, W. R. Hazlewood, and K. Makice. "Twitterspace: A Co-developed Display Using Twitter to Enhance Community Awareness". In: *Proceedings of the Tenth Anniversary Conference on Participatory Design 2008*. Indiana University, 2008, pp. 230–233. ISBN: 978-0-9818561-0-0.

[RN07] M. Riketta and S. Nienaber. "Multiple Identities and Work Motivation: The Role of Perceived Compatibility between Nested Organizational Units*". In: *British Journal of Management* (s1 2007), S61–S77. ISSN: 1467-8551.

[RPP+17] G. Rossetti, L. Pappalardo, D. Pedreschi, and F. Giannotti. "Tiles: an online algorithm for community discovery in dynamic social networks". In: *Machine Learning* (2017), pp. 1213–1241. ISSN: 1573-0565.

[Saa12] D. G. Saari. *Geometry of Voting*. Springer Science & Business Media, 2012. 388 pp. ISBN: 978-3-642-48644-9.

[SC17] H. Shirado and N. A. Christakis. "Locally noisy autonomous agents improve global human coordination in network experiments". In: *Nature* (2017), pp. 370–374. ISSN: 1476-4687.

[SD03] W. R. Scott and G. F. Davis. "Networks in and around organizations". In: *Organizations and Organizing* (2003).

[Shi09] C. Shirky. *Here comes everybody: the power of organizing without organizations*. Penguin Books, 2009. 344 pp. ISBN: 978-0-14-311494-9.

[SKC09] D. A. Shamma, L. Kennedy, and E. F. Churchill. "Tweet the Debates: Understanding Community Annotation of Uncollected Sources". In: *Proceedings of the First SIGMM Workshop on Social Media*. ACM, 2009, pp. 3–10. ISBN: 978-1-60558-759-2.

[SLC+12] H. Sundaram, Y. Lin, M. D. Choudhury, and A. Kelliher. "Understanding Community Dynamics in Online Social Networks: A multidisciplinary review". In: *IEEE Signal Processing Magazine* (2012), pp. 33–40. ISSN: 1053-5888.

[SM00] J. Shi and J. Malik. "Normalized cuts and image segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2000), pp. 888–905. ISSN: 0162-8828.

[SOM10] T. Sakaki, M. Okazaki, and Y. Matsuo. "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors". In: *Proceedings of the 19th International Conference on World Wide Web*. ACM, 2010, pp. 851–860. ISBN: 978-1-60558-799-8.

[STA07] X. Shi, B. Tseng, and L. A. Adamic. "Looking at the Blogosphere Topology through Different Lenses". In: *Ann Arbor* (2007), p. 48109.

[Str01] S. H. Strogatz. "Exploring complex networks". In: *Nature* (2001), pp. 268–276. ISSN: 1476-4687.

[TL10] L. Tang and H. Liu. "Community Detection and Mining in Social Media". In: *Synthesis Lectures on Data Mining and Knowledge Discovery* (2010), pp. 1–137. ISSN: 2151-0067.

[Was94] S. Wasserman. *Advances in Social Network Analysis: Research in the Social and Behavioral Sciences*. SAGE, 1994. 320 pp. ISBN: 978-0-8039-4303-2.

[Wel08] B. Wellman. "The Development of Social Network Analysis: A Study in the Sociology of Science". In: *Contemporary Sociology; Washington* (2008), pp. 221–222. ISSN: 00943061.

[WF94] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994. 852 pp. ISBN: 978-0-521-38707-1.

[WJL+16] W. Wei, K. Joseph, H. Liu, and K. M. Carley. "Exploring characteristics of suspended users and network stability on Twitter". In: *Social Network Analysis and Mining* (2016), p. 51. ISSN: 1869-5469.

[WS98] D. J. Watts and S. H. Strogatz. "Collective dynamics of 'small-world' networks". In: *Nature* (1998), p. 440. ISSN: 1476-4687.

[XKS13] J. Xie, S. Kelley, and B. K. Szymanski. "Overlapping Community Detection in Networks: The State-of-the-art and Comparative Study". In: *ACM Comput. Surv.* (2013), 43:1–43:35. ISSN: 0360-0300.

[YL15] J. Yang and J. Leskovec. "Defining and evaluating network communities based on ground-truth". In: *Knowledge and Information Systems* (2015), pp. 181–213. ISSN: 0219-1377, 0219-3116.

[YLL+14] Y. Yang, C. Lan, X. Li, B. Luo, and J. Huan. "Automatic Social Circle Detection Using Multi-View Clustering". In: *Proceedings of the 23rd ACM CIKM*. ACM, 2014, pp. 1019–1028. ISBN: 978-1-4503-2598-1.