A multilevel clustering technique for community detection

Isa Inuwa-Dutse*, Mark Liptrott, Ioannis Korkontzelos**

Department of Computer Science, Edge Hill University,
St Helens Rd, Ormskirk, L39 4QP, Lancashire, United Kingdom

Abstract

A network is a composition of many communities, i.e., sets of nodes and edges with stronger relationships, with distinct and overlapping properties. Community detection is crucial for various reasons, such as serving as a functional unit of a network that captures local interactions among nodes. Communities come in various forms and types, ranging from biologically to technology-induced ones. As technology-induced communities, social media networks such as Twitter and Facebook connect a myriad of diverse users, leading to a highly connected and dynamic ecosystem. Although many algorithms have been proposed for detecting socially cohesive communities on Twitter, mining and related tasks remain challenging. This study presents a novel detection method based on a scalable framework to identify related communities in a network. We propose a multilevel clustering technique *(MCT)* that leverages structural and textual information to identify local communities termed *microcosms*. Experimental evaluation on benchmark models and datasets demonstrate the efficacy of the approach. This study contributes a new dimension for the detection of cohesive communities in social networks. The approach offers a better understanding and clarity toward describing how low-level communities evolve and behave on Twitter. From an application point of view, identifying such communities can better inform recommendation, among other benefits.

Keywords: Clustering, Multilevel clustering, Community detection, Twitter, Social networks

1. Introduction

A network comprises of many sub-networks or communities with distinct and overlapping properties. Networks exhibit varying degrees of organisations [1], and discovering the structure of various network forms has been investigated [2, 3, 4, 5]. As network size increases, so does the possibility of fragmentation [6, 7], leading to a decrease in the homogeneity of behaviour and attitude across groups [8]. Because similarity breeds attraction and interaction [9], network communities are defined by sets of nodes and edges with strong relationships. Communities are a fundamental organisation principle, especially in vast networks, allowing to analyse the structure and function of networks [3, 10]. Identifying local network structures: (a) provides a means for complex network analysis [11], for applications such as the detection of inter-related web-pages [12, 13], (b) allows to detect cliques [10] and facilitates intelligent recommendations [14] (c)

*Corresponding author
**Principal corresponding author
Email addresses: isadutse91@gmail.com (Isa Inuwa-Dutse), Mark.Liptrott@edgehill.ac.uk (Mark Liptrott), Ioannis.Korkontzelos@edgehill.ac.uk (Ioannis Korkontzelos)

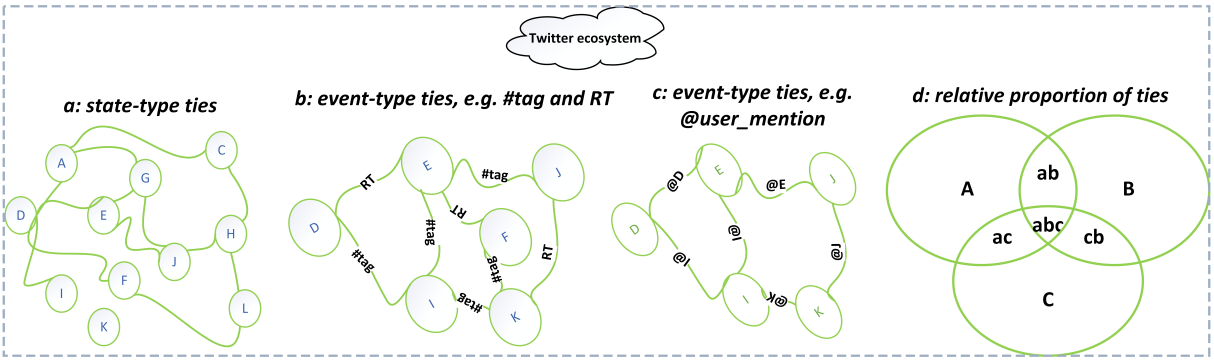

Figure 1: Examples of *event-type ties* ((a),(b) and (c)) on Twitter, allowing users to openly connect via: (a) unidirectional or directed means (e.g. friend or follower), bidirectional or undirected (among friends and followers) (b – c), and indirect or *transitory events*: *retweets, mentions* or *likes*. These flexible connections challenge cohesive community detection and contribute to the proliferation of spurious content. In (d), $A = \{a, a_1, ..., a_l\}, B = \{b, b_1, ..., b_m\}$, and $C = \{c, c_1, ..., c_n\}$ denote users. Reciprocal ties (e.g. ac) or transitive ties (e.g. abc) are rare.

allows to discover organisational principles of networks [15, 16], and (d) helps in studying social behaviour of users [17]. Examples of biological, social or technological networks where community detection has been applied are: *protein–protein interaction networks* [18], *social networks* [4, 10], *food webs* [11], *collaboration networks*, [19] and the *World Wide Web* [4].

The underlying difference across many network communities refers to the definition of connections: some are deterministic, while some are just probabilistic and potentially non-deterministic. Social media, e.g., *Twitter* and *Facebook*, connect a myriad of diverse users, leading to a highly connected, dynamic ecosystem. The complexity and dynamism of this ecosystem results in multiple interaction types at various layers of granularity and intensity: global or local, positive or negative, influential or not, high or low-level. Such interactions culminate in the formation of communities at various levels. Despite the proliferation of various community detection methods, identifying socially cohesive communities on Twitter still poses challenges. Communities with low presence are implicit and require extensive exploration to understand the mechanism governing their behaviour [20]. Since social networks exhibit properties from other networks [10], the limitations of existing approaches are due to:

Methodological viewpoints and connection types. Social network theorists hold two viewpoints in investigating social relationships in a network: realist, based on a pre-conceived notion of the existence of relationships, and nominalist, based on questions posed by the investigator [21]. Moreover, social ties are formed around *event-type ties*, a transitory connection that often results in socially distant members. Such connections on Twitter include subscribing to trendy hashtags or retweeting popular users. *State-type ties* are based on static, structural connections among users, which suggest familiarity and trust [22]. Community detection on Twitter focusses mostly on directed connections (*event-type ties*) based on the *realist's* approach. This is valid in many networks, but can lead to many unrelated sets of users. We argue that the wealth of connection forms on Twitter, shown in Figure 1, contribute to widespread spurious content and imply the existence of less cohesive user communities.

Proliferation and complexity of online content. a rapid increase in network size increases the likelihood of fragmentation [6, 7], which in turn decreases the homogeneity of behaviour and

attitude across groups [8]. With an average 139m daily users contributing to 500m content[1], it is becoming more challenging to keep track of socially cohesive communities on Twitter. Furthermore, large scale and transitory content (mostly from influential users) often dominate the space leading to many forms of explicit communities [23]. Thus, basing a community detection task on transitory aspects of *metadata* such as popular hashtags or trending topics does not often reflect true connectivity [24], hence limiting the full realisation of the benefits in communities such as *cliquishness* and *local coherence*.
This study attempts to address the identified challenges, to advance our knowledge concerning community detection problems.

1.1. A Multilevel Clustering Technique

A community detection paradigm involves prediction and quantification to identify a community structure and relevant details about a network [25]. Predicting membership and assigning items to clusters is achieved using equivalence measures or scoring functions. Establishing the equivalence of network entities is achieved based on (a) equivalent units with the same connection pattern to the same neighbours or (b) equivalent units that have the same or a similar connection pattern to different neighbours [26]. Accordingly, communities are formed around two primary modalities or information sources: *network structure* and *node attributes*. Until recently, community detection methods relied on a single information source. Conventional methods such as *normalised cut* [27] and modularity [28] rely on the topological structure of networks. A bi-modal approach, based on network structure and the corresponding features or attributes of nodes as information sources, is becoming popular [29, 30, 31, 16]. According to Figure 1, connections on *Twitter* may manifest differently, such as *sharing a link*, *re-tweeting (RT)*, using the same or similar *hashtags*, *user mention (@)* or *follower-ship*. Such connections are porous, allowing to connect with many diverse users and hindering the identification of cohesive groups. Noting that these eccentric connections patterns can mislead the detection of socially related users and encourage the propagation of spurious content, we propose a *multilevel clustering technique (MCT)* to identify socially cohesive user groups on Twitter, termed *microcosms*. No practical reasons prevent MCT to apply to other domains that involve network data. However, it would require minor amendments for platforms where a reciprocal tie is the default connection, e.g. Facebook. Failing to recognise Twitter's eccentric topological structure would make the approach less generalisable. Focusing on Twitter leads to a more encompassing framework that can be mapped to other networks.

MCT measures similarity within a community of users using local and global information, modelling *structural* and intrinsic *textual* features jointly. In Figure 2(a) and (b) user communities exist based on *structural* and *content* or *textual* similarity, respectively. Users under the structural component, a form of a *state-type tie*, are related based on reciprocal ties, which are rare in Twitter, and the community is more cohesive than the community of users based on content or textual similarity, a form of an *event-type tie*. A more cohesive community is the one that recognises both structural and content similarity, in Figure 2(c). Intuitively, the degree of cohesiveness varies across different communities: a community based on both modalities is the most cohesive, followed by a community based on high structural similarity but low or no content or textual similarity. Finally, the least cohesive community exhibits high similarity in the

[1] See www.omnicoreagency.com/twitter-statistics

Table 1: Relevant notations and their corresponding descriptions

Notation	Description
$\mathcal{D}$	network data
$\mathcal{V}$ and $\mathcal{E}$	sets of nodes and edges, respectively
m_{v_i}	denotes the network of a user v_i
fr_{v_i} and fl_{v_i}	denote sets of friends and followers of user v_i
$\kappa \in m_{v_i}$	set of reciprocal ties of user v_i
$\mathcal{A}_f$	set of all possible attributes of a node
$\mathcal{X}_f$	set of features inducing reciprocity
$a \succ b$	a binary relation between a and b
$\exp x$ or e^x	base of a natural logarithm, ln, $\exp(lnx) = x$; $e \approx 2.71828$
d_i, p_i, q_i, u_i	respective ith component of vectors, $\mathbf{D}, \mathbf{P}, \mathbf{Q}, \mathbf{U}$
τ	a predefined threshold for comparison, e.g. $\tau \geq 0.5$
$\mathcal{S}_r$ and $\mathcal{T}_r$	sets of structurally-related and textually-related nodes, respectively

textual component but low or no structural similarity. Groups of structurally similar nodes are analysed by *spectral clustering*, which involves a series of methods ranging from adjacency and affinity matrices to dimensionality reduction. The textual component complements the structural aspect through a form of document-pivot clustering, in which weights are assigned to features in the document according to a weighing scheme [32, 33, 34, 35].

1.2. Contributions

MCT relies on reciprocal ties, based on the assumption that combining structural and textual features offers a more cohesive community representation. Our contributions are two-fold:

A new dimension to the detection of cohesive communities. The ability to follow anyone on Twitter results in many unidirectional connections between socially unrelated users, affecting clustering and the integrity of online content. To counter the challenging and time-consuming task of collecting large scale reciprocal ties on Twitter, we proposed a strategy that returns the likelihood of reciprocity among users. As a result, the detection of socially cohesive communities is enhanced, providing a useful analysis tool and strengthening the validity of online content. Moreover, by identifying communities of users with a strong cohesion, a well-informed recommendation that recognises structural and textual similarities can be achieved.

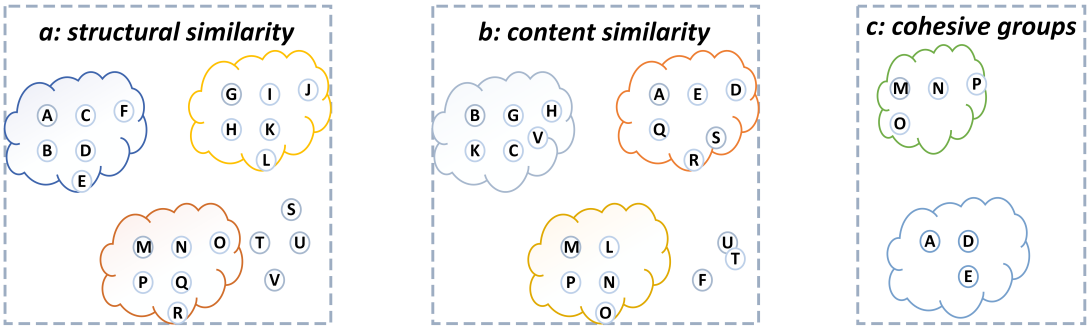

Figure 2: Node clustering in MCT in three stages, according to (a) structural similarity, (b) *content* or *textual* aspect similarity, and (c) joint *structural* and *content* similarity.

A bi-modal community detection approach. MCT addresses the problem of structurally unrelated users by adding a layer of social cohesion to existing community detection methods. Specifically, MCT advances existing techniques through: (a) an in-depth utilisation of a bi-modal source of information for community detection, (b) detection of network communities at various levels, (c) a robust and scalable community detection algorithm that is less prone to noise in the network data, and (d) an intuitive interpretation of the detected communities.

The remaining of this paper is structured as follows. Section 2 provides background details and related work. Section 3 formulates the problem and describes the MCT framework. We describe the experimentation process in Section 4 and discuss the main observations in Section 5. Finally, Section 6 concludes the study and provides some considerations for future work. For ease of referencing, Table 1 summarises the notations used.

2. Background

Humans can effortlessly abstract complex phenomena, but efficiently automating the process is daunting partly due to the multidimensional nature of clustering data [36]. In this section, we review relevant topics and studies associated with clustering and community detection tasks.

2.1. Network and community structure

A network comprises of heterogeneous nodes connected via edges. The topological structure of networks and other quantities related to them are useful in understanding complex networks across numerous domains [4]. Various levels of relationship forms in networks have been analysed, from the structure of microscopic organisms to complex networks, such as the internet [2, 3, 4]. Complex networks were once considered to as random and the classic random graph model [37] was the standard analysis tool until regular patterns in various networks were discovered, e.g., via statistical analysis. Fundamentally, network complexity [38] is defined by: (a) *Clustering coefficient* quantifies the probability of a node to be clustered, assuming that users with common friends are likely to know each other. (b) *Degree distribution* quantifies the probabilistic distribution of nodes. (c) *Small-worldliness* is a network property associated with short path-length, i.e., many structured short-range connections and few random long-range ones, and network diameter that is exponentially less than its size [3].

In *social media*, communication happen at various layers of granularity and intensity: global or local, positive or negative, influential or not. In contrast with the early unidirectional *two-step* communication model, where few users serve as intermediaries between mass communication and the public [39], the design of social media allows users to generate and consume information. On social media, communication follows the influence network model, enabling multi-way flow, where users can simultaneously generate and consume information [38]. Twitter is dominated by influential users, logically dividing a clique of content pushers and consumers, resembling the two-step flow model [39]. This division is strengthened by strategies, such as content promotion, that entice users to engage more, and to follow or add friends. Using these strategies social media users can increase their network of friends, generating more value to the platforms. A social media network is the synthesis of many user communities, and identifying these structures is a vital task. Because members of a community are highly similar to each other and less or not at all similar to members of other communities [40, 1, 41], a *community structure* has densely connected node groups and sparser connections to other communities [42]. Thus, community

identification involves prediction and quantification tasks to detect the relevant structures and their characteristics [25]. Selecting an effective similarity measure is crucial as it is allows a clustering algorithm to identify groups and affects the *signal-to-noise-ratio* within the instance matrix [43].

2.2. Related work

Network partitioning has attracted interest from various domains of expertise, hence diverse strategies have been put forward to identify relevant communities embedded in a network.

Clustering and Community Detection. Often clustering and community detection are used interchangeably in the literature. Clustering mostly focuses on a single modality, e.g., using node attributes to group network objects, whereas community detection focuses on network structure as a function of connectivity involving social interaction. As a form of dimensionality reduction, clustering entails unsupervised partitioning a network into groups of related objects using a domain-specific scoring function and maximising in-group similarity. There are two principal lines of research in this direction: graph partitioning and hierarchical modelling [41]. We follow this classification, as it reflects the approach in this study. Methods can also be divided in dimensionality reduction based ones, and graph partitioning (hierarchical or not [44]) and hierarchical ones [45].

2.2.1. Graph-based and hierarchical methods

Graph-based clustering assumes that a community structure exists in the network and attempts to discover it using specific techniques. Graph partitioning divides the network into predefined node groups and suits applications where the number and size of groups are known, e.g., in parallelisation of computing processors. The approach may involve hierarchical agglomerative clustering [46] following a random walk model [37], or based on modularity [42] optimisation, such as in the Louvain detection algorithm [47]. The clustering method can be based on *iterative bisection*, which divides the network optimally into two parts and repeats until the required number of partitions is reached [28]. The modularity measure measures community strength and detects groups, assuming that community structures correspond to an interesting statistical arrangement of edges. Positive values indicate the presence of community structures, i.e., that nodes within a community are more tightly connected than by chance [28]. The modularity value of real networks ranges from .3 to .7. The higher the score, the more cohesive the community structure [41]. Predefining the maximum bisection size is required, which may affect performance. Metrics such as *betweenness* or *shortest loop edges*, are central to the operation of algorithms that process graphs to detect groups of similar nodes [48].

Hierarchical modelling follows a different technical approach from graph-based clustering. Assuming that there are natural subgroups in a network, hierarchical clustering utilises a similarity measure, such as Euclidean distance or Pearson correlation to analyse the network [2]. In particular, pairwise node similarities are computed and nodes are iteratively and deterministically assigned to clusters. Commonly, similar clusters are iteratively merged into larger ones [45]. Furthermore, categorisation based on a generative or model-based and discriminative or similarity-based is used in the literature. Model-based or generative clustering algorithms, e.g., Latent Dirichlet Allocation (LDA) [49, 29, 50], are a form of Expectation Maximisation (EM) that aim to learn a generative model from data segments, where each model represents a cluster [51]. The EM-based models estimate the maximum likelihood of data-points to belong

to a cluster and is suitable for incomplete data. On the other hand, similarity-based clustering algorithms are based on optimising a scoring function that is used to compute pairwise similarity between data-points. This form of clustering follows hierarchical agglomerative clustering or block modelling.

2.2.2. *Multiview and bi-modal clustering*

Multiview and bi-modal techniques aim to improve clustering performance using multiple independent data sources; thus, multiview clustering relies on data that can be split into independent sub-features or attributes [52, 53]. For instance, a *web page* can be described by its textual content and pages that link to it [54, 55]. The advantages of multiview clustering over its single view counterpart has been investigated using algorithms based on *K-means* and *Expectation Maximisation* [54].

Bi-modal clustering technique is based on the fact that network communities are formed around two primary modalities or information sources: *network structure* and node *attributes*. In many cases, structural and textual aspects evolve simultaneously and communities are discovered according to the nodes' similarity degree vis-á-vis those two aspects. Until recently, studies mostly focused on one aspect, not both [16, 29, 31, 56, 57]. A study closely related to our approach, proposes a generative model for networks with node attributes [16]. However, the depth of the features, especially the nodes' attributes, is shallow and the node attribute (*hashtag*) is insufficient to analyse the depth of similarity between network entities in a complex environment such as Twitter in which the structural component is not fully captured due to reliance on directed edges. The connected k-centre approach employs both structural and attribute information for a given network partition [56]. The problem is NP-hard, leading to many heuristics. Similarly, SA-cluster method combines structural and attributes' similarities for community detection by partitioning a network into cohesive k-clusters with structural and attribute information using a distance metric to estimate pairwise node similarity or closeness [57].

Conventional methods, such as normalised cut [27] and modularity [28], are based on topological structures. However, many networks come with incomplete information, e.g., a *terrorist network* or food web [30]; thus, community detection in networks with edge uncertainty or incomplete information is getting traction. Inferring links in incomplete networks is challenging, because the information is usually localised within a small, linked group. The full wealth of data has been used to learn a generalisable distance metric to complete the missing information [30]. However, this approach is too complicated and does not account for the breadth required in *textual* aspects in networks with many transient connections, such as Twitter[2]. The MCT is a two-stage clustering technique that recognises different modalities as information sources; it incorporates multiview aspects at various levels, structural and textual, using independent features.

3. MCT framework

Noting that nodes in a community are highly similar and edges among communities are infrequent, community detection is usually formalised to identify network partitions that satisfy specific requirements. The problem focuses on detecting smaller groups with high similarities, using a joint similarity function that considers global and local information as a two-stage process

[2]As shown in Figure 1, Twitter communities are formed based on many factors.

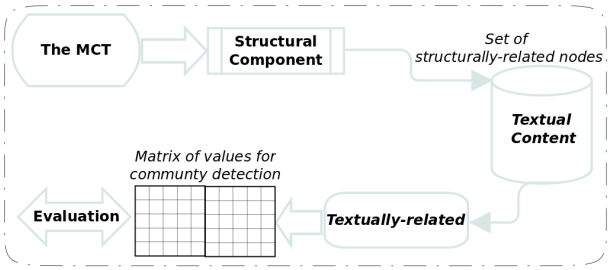

Figure 3: Execution pipeline of the MCT framework - The structural component processes a collection of structurally similar nodes, promoting group formation among them. Then, *textual analysis* of structurally related node content identifies groups according to discussion topics. MCT combines *state-type* and *event-type* ties.

comprising of structural and textual components (shown in Figure 2). Figure 3 shows a block diagram of the stages in the MCT framework.

3.1. Structural component

The structural component is based on dyadic, pairwise edge between two users, and transitive ties, which are the basic forms of establishing reciprocal ties in social networks. We aim to identify groups of users with true reciprocal relationships at dyadic and transitive level. Transitivity expresses the social preference to be friends with a *friend-of-a-friend* and has been characterised as a peculiar network feature [3]. Transitive ties are synonymous to *Simmelian ties*, strong social relationships among three or more individuals, which are vital in understanding a network's social tie structure [58]. Our approach assumes that community detection or clustering methods that take reciprocal ties into account offer a more cohesive community representation. Our analysis of Twitter datasets concluded that true reciprocal ties are rare. Thus, we use a method that strengthens the possibility of finding Twitter users with reciprocal ties. A user with many reciprocated ties can represent a *microcosms*, allowing to analyse a user group as a unit. Research in social science suggest that users compare themselves with one another and adopt similar behaviour with users similar to them [9]. Homophily on Twitter can be interpreted as a reciprocal relationship among users. Noting this insight and the inspiration drawn from social homophily, we argue that users with similar profiles are more likely to connect on Twitter. Therefore, structural equivalence is mapped to a *state-type tie* to infer structural similarity according to the node's attributes. Figure 4 shows features that contribute in finding structurally related nodes.

3.1.1. Modelling structural clusters

Definitions. This section begins with the definitions of relevant concepts and terms in the implementation. Table 1 provides a summary of all relevant notations used in the study.

- *Network data* $\mathcal{D}$ consists of sets of nodes, $v_1, v_2, ..., v_m \in \mathcal{V}$, and edges, $e_1, e_2, ..., e_n \in \mathcal{E}$. Each node is described by its structural and textual features, as shown in Figure 4.

- *Dyadic and transitive ties:* a relation, $\succ$, between two nodes $v_i, v_j \in \mathcal{D}$ is dyadic[3] if v_i follows v_j and vice versa, i.e. $v_i \succ v_j \Leftrightarrow \forall v_i, v_j \in \mathcal{D}$. In this context, v_i follows v_j is a directed relationship; if v_j follows v_i back, it is undirected; see some examples in Figure 5.

[3]Dyadic tie, pairwise, 2-star or binary relations are used interchangeable in this study.

A transitive or *Simmelian tie* is a social relationship within groups of three or more. A binary relation, $\succ$, over a set $\mathcal{D}$ is *transitive*: $v_i \succ v_j, v_j \succ v_k \Leftrightarrow v_i \succ v_k, \forall v_i, v_j, v_k \in \mathcal{D}$.

- *Prediction features and reciprocity:* To identify structurally related nodes, we use features, such as *indegree (ind)*, the number of incoming edges to a node; *outdegree (out)*, the number of outgoing edges from a node; and *category (cat)*, indicating if a node is *verified* or *unverified*. Account verification ascertains legitimacy of the account holder. $\mathcal{A}_f$ denotes the set of all possible node features, which can be used to extract feature subsets, $\mathcal{X}_f = \{ind, out, cat\} \subset \mathcal{A}_f$, from a profile, as in Figure 4. The features of a node pair, v_i, v_j, are denoted by: $\mathcal{X}_{f_{v_i}} = \{ind_{v_i}, out_{v_i}, cat_{v_i}\}$ and $\mathcal{X}_{f_{v_j}} = \{ind_{v_j}, out_{v_j}, cat_{v_j}\}$ and are used for training models that predict the likelihood of *reciprocity*.

Nodes reciprocity and constant error. We predict node sets with dyadic ties on Twitter using the approach in [59], and we use them for clustering. The formulation assumes that reciprocity is based on the features that can induce friendship, in Figure 4(b). Nodes reciprocity hypothesises that the decision to reciprocate or establish friendship correlates with the idea of homophily and structural equivalence. Reciprocal ties are predicted based on these concepts between node pairs.

Consider the sets of nodes, $\mathcal{V}$, and edges, $\mathcal{E}$. The likelihood of *reciprocity*, $p(R_{v_i,v_j}), \forall v_i, v_j \in \mathcal{V}$) involved in the computation of reciprocal units (see Section 4.2.1) is described by Eq. 1 to 3, leading to *reciprocal-communities* $\mathcal{C}_{rc}$. The features of a node pair, v_i, v_j, for comparison are denoted by: $\mathcal{X}_{f_{v_i}}$ and $\mathcal{X}_{f_{v_j}}$, such that the ratio of the attributes, e.g. *ind* or *out*, is a real value quantity given by: $\frac{ind_{v_i}}{ind_{v_j}} \in \mathbb{R} \;\; \forall f_r \in \mathcal{X}_{f_{v_i,v_j}}$. If the comparison evaluates to a value in $[0.75, 1.25]$, the pairs are assumed to have similar attributes (1), otherwise dissimilar attributes (0). The interval is to allow extra freedom for minor discrepancies between the features. For instance, if the ratio equals 1.0, the pair has identical attributes, which is useful in analysing aspects of homophily and social equivalence. The binary feature comparison values are used to

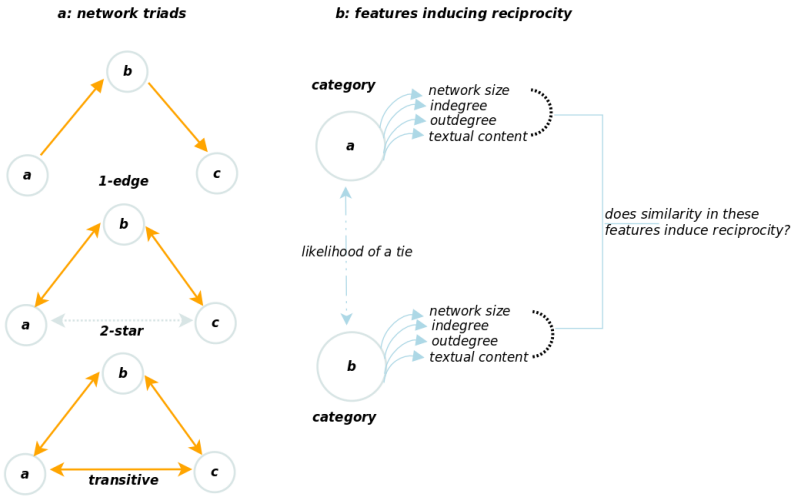

Figure 4: (a) Possible social ties in a network triad - Each node is associated to a set of nodes with a directed or reciprocal tie. (b) An example dyad and the features responsible for tie formation between users on Twitter. A probability score is assigned to each feature, to discover their inter-dependencies and enable reciprocal ties.

compute the overall similarity using the *Jaccard Index, J*:

$$J(\mathcal{X}_{f_{v_i}}, \mathcal{X}_{f_{v_j}}) = \frac{|\mathcal{X}_{f_{v_i}} \cap \mathcal{X}_{f_{v_j}}|}{|\mathcal{X}_{f_{v_i}} \cup \mathcal{X}_{f_{v_j}}|} \tag{1}$$

Moreover, modelling reciprocity or the response to a *friendship* request is associated with a decision error, that can be quantified based on *response probability*. Response probability applied to cases where a feature set for a decision involves a constant probability of error (Eq. 2) in the choice [60]. Thus, the probability of *reciprocity* between pairs based on the similarities in their features $J(v_i, v_j)$ (Eq. 1) is given by [61]:

$$\epsilon_{v_i,v_j} = \frac{1}{\zeta \times (1 + \log(J(v_i, v_j) + \zeta))} \tag{2}$$

The *constant error term*, ζ, is assigned the value of $1/3$ and the final relation is given by:

$$p(R_{v_i,v_j}) = \frac{1}{1 + \exp \varphi} \tag{3}$$

where $\varphi = -\log(\epsilon_{v_i,v_j} + J(v_i, v_j)) \times (\epsilon_{v_i,v_j} + J(v_i, v_j))$. Any node pair v_i, v_j are structurally similar or related $\mathcal{S}_r$, if their degree of reciprocity (Eq. 3) is greater than a predefined threshold τ. It follows that $\mathcal{S}_r : \forall v_i \in \mathcal{S}_r \; \exists v_j$, such that $p(R_{v_i,v_j}) \geq \tau$, where $\mathcal{S}_r \subset \mathcal{D}$.

Collection of structurally related nodes. Eq. 3 allows to identify as many node sets with a high likelihood of establishing reciprocal ties, thus adding a layer of social cohesion to the MCT framework. Identifying groups of structurally related nodes begins with a high-level aggregation of nodes according to network size (for *network-communities*) and reciprocity (for *reciprocal-communities*). This led to the question *what does it mean for nodes to be structurally related?* As a simplistic example, consider a finite set $\mathcal{V}_{13}$ that contains 13 nodes: $\{v_1, \cdots, v_{13}\} \in \mathcal{V}_{13}$. After executing *Algorithm f-sim* (Algorithm 1), which predicts the likelihood of reciprocity, the following pairs of nodes are structurally-similar or related[4]: $v_1 \sim v_2, v_1 \sim v_3, v_1 \sim v_5, v_2 \sim v_4, v_2 \sim v_5, v_2 \sim v_9, v_3 \sim v_{11}$. Accordingly, three structurally related communities can be identified: $c_1 = \{v_1, v_2, v_3, v_5\}$, $c_2 = \{v_2, v_4, v_5, v_9\}$ and $c_3 = \{v_3, v_9\}$. The communities can be expressed in a matrix form for spectral analysis. Matrix entries can be based on states, such as the *reciprocity potential* of nodes defined as the ratio of *outdegree* over *network size*.

Spectral analysis. Since structurally related nodes can be easily transformed into a graph, we apply *spectral analysis* to identify clusters and enabling *Sociometry*[5], a means to measure social relationships [62]. Spectral clustering involves operations ranging from the construction of

[4]The symbol '$\sim$' is used to denote structural similarity between pairs.

[5]Metrics such as structural hole, homophily, and centrality metrics – degree, closeness, betweenness

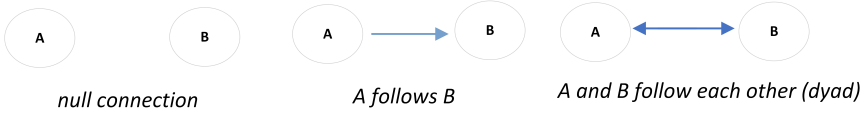

Figure 5: Examples of possible relationships between pairs

Algorithm 1 *: Algorithm f-sim* returns the likelihood of reciprocity between pairs.

Initialisation: $\{\} \longleftarrow \mathcal{S}_r; \{\} \longleftarrow \mathcal{S}_u$
Input: a finite collection of network data $\mathcal{D}$
while $\mathcal{D} \neq \emptyset$ **do**
    $\forall v_i, v_j \in \mathcal{D}$, compute $p(R_{v_i,v_j})$ using Eq. 3 ▷ $v_i \neq v_j$
    if $p(R_{v_i,v_j}) \geq \tau$ **then** ▷ τ, *a predefined threshold*
        $\mathcal{S}_r \leftarrow (v_i, v_j)$ ▷ *structurally related*
    else
        $\mathcal{S}_u \leftarrow (v_i, v_j)$ ▷ *structurally unrelated*
    end if
end while
Output:
    $\mathcal{S}_r, \mathcal{S}_u, \mathcal{M}_{A_{i,j}}^{n \times n}$ ▷ *adjacency matrix* $\mathcal{M}_{A_{i,j}}^{n \times n}$

adjacency or affinity matrices to clustering in a reduced dimension [63]. We construct the *adjacency*, *similarity* and *degree* matrices based on ground-truth data and Eq. 3. The adjacency matrix $\mathcal{M}_{A_{i,j}}$ (Eq. 4) encodes the structural similarities among node pairs. Presence or absence of reciprocity is marked with 1 or 0, respectively.

$$\mathcal{M}_{A_{i,j}} = \begin{cases} 1 & \text{if } p(R_{v_i,v_j}) \geq \tau \\ 0 & \text{if } otherwise \end{cases} \tag{4}$$

The *degree matrix*, $\mathcal{M}_D$, is a diagonal matrix obtained by summing the entries in Eq. 4 across the rows, in which entry i, i denotes the degree or number of edges to node i. Thus, each entry in the diagonal d_i (Eq. 5), of matrix $\mathcal{M}_D$ is defined by:

$$d_i = \sum_{\{j|(i,j) \in E\}} p(R_{v_i,v_j}) \geq \tau \tag{5}$$

Subtracting the adjacency matrix, $\mathcal{M}_{A_{i,j}}$, from the degree matrix, $\mathcal{M}_D$, gives the graph *Laplacian*, $\mathcal{M}_L = \mathcal{M}_D - \mathcal{M}_A$, whose *eigenvectors* and *eigenvalues* offer informative features for clustering. Diagonal entries in Eq. 6 denote the degree of nodes, off-diagonal negative entries ($-p(R_{v_i,v_j})$) represent edges between node pairs and zero entries signify no edges.

$$\mathcal{M}_{L_{i,j}} = \begin{cases} d_i & \text{if } i = j \\ -p(R_{v_i,v_j}) & \text{if } p(R_{v_i,v_j}) \geq \tau \quad \triangleright\, edge(i,j)\ exists \\ 0 & \text{if } otherwise \end{cases} \tag{6}$$

3.1.2. Structural optimisation

Given a set of structurally related nodes, hidden local communities can be uncovered via matrix decomposition on the following matrices of interactions and corresponding dimensions:

- $\mathcal{M}_{c_{ns}} \mapsto n \times p$: a matrix of nodes according to a concept, e.g., network size
- $\mathcal{M}_{c_{vr}} \mapsto n \times k$: a matrix of nodes according to reciprocity
- $\mathcal{M}_{c_{nr}} \mapsto p \times k$: a matrix of high-level and local communities

The network-communities matrix, $\mathcal{M}_{C_{ns}}^{n\times p}$, is decomposed into its approximate constituents:

$$\mathcal{M}_{C_{ns}} \approx \mathcal{M}_{vr}\mathcal{M}_{C_{nr}}^{T} \tag{7}$$

Interpretability is desirable in the MCT framework, hence, the decomposition follows a non-negative matrix factorisation (NMF) scheme [64]. NMF provides an intuitive factorisation, in which non-negative constraints are imposed on the optimisation parameters [65].

$$min._{\mathcal{M}_{c_{vr}},\mathcal{M}_{c_{nr}}} ||\mathcal{M}_{c_{ns}} - \mathcal{M}_{c_{vr}}\mathcal{M}_{c_{nr}}^{T}||_F^2, \text{ subject to } \mathcal{M}_{c_{vr}}, \mathcal{M}_{c_{nr}} \geq 0 \tag{8}$$

We simplify the notation of matrices as follows: $\mathcal{M}_{c_{ns}} \mapsto D$, $\mathcal{M}_{c_{vr}} \mapsto P = [p_{is}]$, $\mathcal{M}_{c_{nr}} \mapsto Q = [q_{js}]$. The formulation in Eq. 8 allows to consider *Lagrangian relaxation* to optimise the squared Frobenius norm ($||\cdot||_F^2$) of the matrix. Consequently, NMF's non-negative constraints are relaxed by introducing the *Lagrangian multipliers*, two new parameters (α, $\beta \leq 0$), to the corresponding entries of the optimisation parameters (P, Q). Accordingly, the objective function M_{sr} is expressed as a *minmax* problem, that requires a simultaneous minimisation over P, Q and maximisation[6] over all applicable values of α and β:

$$M_{sr} = ||D - PQ^T||_F^2 + \sum_{i=1}^{n}\sum_{s=1}^{k} p_{is}\alpha_{is} + \sum_{j=1}^{p}\sum_{s=1}^{k} q_{js}\beta_{js} \tag{9}$$

The optimisation starts by computing the gradient of the relaxed Lagrangian, with respect to the first aspect of the *minmax* (i.e. minimisation) optimisation variables. Although α and β offers a degree of flexibility (that comes at a cost), optimal solution requires optimisation conditions to be based on P, Q only. We apply a handy technique based on Karush-Kuhn-Tucker (KKT) optimality condition [66] to eliminate the Lagrangian multipliers. The KKT condition suggests that $p_{is}\alpha_{is} = q_{js}\beta_{js} = 0$. Non-negative random values in $(0, 1]$ are iteratively assigned to the parameters P and Q based on the following update rule (after simplifying equations as shown in the Appendix):

$$p_{is} \Leftarrow \frac{(DQ)_{is}p_{is}}{(PQ^TQ)_{is}},\ q_{js} \Leftarrow \frac{(D^TP)_{js}q_{js}}{(QP^TP)_{js}},\ \forall i \in \{1,\cdots,n\}, j \in \{1,\cdots,n\}, s \in \{1,\cdots,k\} \tag{10}$$

The process of updating P and Q involves comparing their values to the original matrix D, aiming in minimising the difference or *error*. Parameters p_{is} and q_{js} are iteratively updated until convergence. A successful iterative update process ensures that the underlying matrices exhibit strong correlations among their respective entries.

3.2. Textual component

The textual component applies a form of document-pivot clustering based on weighted features [32, 33, 34, 35]. Due to the volume of tweets and their short length, it is difficult to gain a broad perspective about topics. Hence, using a single tweet may not provide sufficient information. To understand the discussion topics and the degree of similarity among tweets, a fixed number of tweets is extracted from nodes in the structurally related sets, $\mathcal{S}_r$. We

[6]This is needed because the Lagrangian multipliers are initialised with negative values.

utilise Latent Dirichlet Allocation (LDA), which has been previously applied for similar tasks [67, 50, 68]. LDA is a probabilistic generative model that assigns word distributions to *topics* and topic distributions to *documents* in a corpus, so that the documents represent random mixtures over latent topics [49]. In this study, tweets collected from each node $v_i \in \mathcal{S}_r$ define a corpus $\mathcal{T}_{v_i}$, whose overall theme is analysed for comparison with other nodes.

3.2.1. Modelling textual clusters

Identifying textually related nodes, $\mathcal{T}_r$, starts by aggregating a finite collection of textual content, $\mathcal{T}$, from each node v_i. The collection of k tweets produced by node v_i over time, i.e. $\{t_{i1}, t_{i2}, t_{i3}, ..., t_{ik}\} \in \mathcal{T}_{v_i}$ consists of a set of m n-gram features $\{f_{i1}, f_{i2}, f_{i3}, ..., f_{im}\} \in t_i \in \mathcal{T}_{v_i}$. Each stream of tweets is preprocessed to extract shingles[7] for transformation following the *term-frequency-inverse document frequency (tf-idf)* weighting scheme [69]. The *tf-idf* vector of a tweet, $\mathbf{v_i}$, can be normalised or not; we apply the $L_2 - norm$ given by: $\mathbf{v_i} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{v_i}}{||\mathbf{v_i}||_2}$.

Similar documents. The collections of tweets for any node pair are: $\mathcal{T}_{v_i}, \mathcal{T}_{v_j} \in \mathcal{T}_r \subset \mathcal{S}_r$. The aggregation scheme ensures that each node has a distinct fingerprint for comparison with others. We train the LDA model so that each tweets in the corpus has a finite distribution over topics, and topics have distributions over words. The distribution of each tweet, dubbed *anchor tweet*, is used to compare with other tweets to locate the most similar tweets and generate relevant matrices. Because LDA-based comparison relies on the probability distributions of tweets, we apply Jensen-Shannon Divergence (JSD), a useful statistical metric, to measure tweet similarity as the degree of divergence in the respective distributions. Unlike Kullback-Leibler Divergence, *JSD* is symmetric, which is crucial in the task of comparing tweets, since similarity should be the same irrespective of the order, i.e.,$X \mapsto Y$ or $Y \mapsto X$ be equal. For example, given two discrete distributions X and Y, *JSD* is defined as:

$$JSD(X||Y) = \frac{1}{2}D(X||\mu) + \frac{1}{2}D(Y||\mu) \tag{11}$$

The *JSD* distance measure (JS_{dist}) is obtained by squaring its *divergence relation*:

$$JS_{dist} = \sqrt{(JSD(X||Y)} \tag{12}$$

It follows that any pairs of tweets, t_i and t_j, are textually-similar or related, $\mathcal{T}_r$, if their similarity degree, ϕ, is greater than a predefined threshold τ. Thus, $\forall t_i \in \mathcal{T}_r \exists t_j : \phi(t_i, t_j) \geq \tau, \;\; \mathcal{T}_r \in \mathcal{S}_r$. Because a finite collection of tweets is extracted from each node in $\mathcal{S}_r$, each $t_i \in \mathcal{T}_r$ consists of a node and its set of tweets. LDA outputs come in a dense $d \times t$ matrix $\mathcal{M}_{lda}^{d \times t}$, consisting of d tweets and their corresponding t topics. Moreover, two matrices apply to the textual component: (a) $\mathcal{M}_{vt} \mapsto m \times q$: matrix of m nodes and top q topics, and (b) $\mathcal{M}_{va} \mapsto m \times m$: affinity matrix of nodes according to topic similarity. Consequently, node communities are formed around common discussion topics and the goal to cluster them according to topical similarities, as in $tr(\mathcal{M}_{va}\mathcal{M}_{vt}\mathcal{M}_{vt}^T)$. Algorithm 2 describes how to obtain the textually related clusters.

[7] *Shingles* are obtained by removing stopwords and other non-content bearing terms in a tweet.

Algorithm 2 *: Algorithm text-sim* identifies *textually related clusters*

Initialisation: $\{\} \longleftarrow \mathcal{T}_r; \{\} \longleftarrow \mathcal{T}_u$
Input: collection of *structurally related nodes* $\mathcal{S}_r$
$\forall v_i \in \mathcal{S}_r$, get k texts $g(\mathcal{T}_{v_i})$ ▷ $g(\mathcal{T}_{v_i})$ *set of k texts of* v_i
  $\mathbf{x_i} \longleftarrow t_i \in g(\mathcal{T}_{v_i})$ ▷ *get texts vectors* $\mathbf{x_i}$
  truncate $\mathbf{x_i}$ ▷ *retain b top terms in vector* $\mathbf{x_i}$
  $m(\mathcal{T}_{v_i}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{x_i}}{||\mathbf{x_i}||_2}$ ▷ *mean of* L_2 *normalised* $\mathbf{x_i}$
  $LDA(m(\mathcal{T}_{v_i}))$ ▷ *invoke the LDA on* $m(\mathcal{T}_{v_i})$
  $\mathcal{T}_{sim}(\mathcal{T}_{v_i}, \mathcal{T}_{v_j}) = JS_{dist}(\mathcal{T}_{v_i}||\mathcal{T}_{v_j})$ ▷ *get similar texts using Eq. 12*
if $\mathcal{T}_{sim}(\mathcal{T}_{v_i}, \mathcal{T}_{v_j}) \geq \tau$ **then**
  update $\mathcal{T}_r$ ▷ *textually related*
else
  update $\mathcal{T}_u$ ▷ *textually-unrelated*
end if
Output:
  $\mathcal{T}_r, \mathcal{T}_u, \mathcal{M}_{ta}^{m\times m}$ ▷ *affinity matrix* $\mathcal{M}_{ta}^{m\times m}$

3.3. Microcosm detection algorithm

The problem of discovering community structures is modelled as a multilevel clustering task, in which nodes are grouped according to scoring functions. Using the affinity matrix based on $\mathcal{S}_r$ or $\mathcal{T}_r$, various algorithms can be used to identify relevant partitions by optimising the separate and joint similarities of $\mathcal{S}_r$ and $\mathcal{T}_r$: $\psi(S_r, T_r) = \phi(S_r) + \phi(T_r)$. A cohesive community is a collection of nodes, $\mathcal{V}$, with high degree of similarity structurally and textually. Thus, the microcosm detection problem can be formally defined as:

> given a collection of network data $\mathcal{D}$, defined by sets of nodes, $\mathcal{V}$ and edges, $\mathcal{E}$, for each node $v_i \in \mathcal{V}$ consisting of sets of *structural* and *textual* features, the goal is to identify a collection of highly cohesive sub-networks $\mathcal{P}$:
>
> $$\mathcal{P} : \forall v_i \in \mathcal{S}_r \; \exists v_j : p(R_{v_i,v_j}) \geq \tau \text{ and } \forall t_i \in \mathcal{T}_r \subset \mathcal{S}_r \; \exists t_j : \phi(t_i, t_j) \geq \tau, \;\; \mathcal{P} \subset \mathcal{D}$$

The above formulation means that for all node pair in the partition $\mathcal{P}$, both the *structural* and *textual* similarities are greater than their respective threshold τ[8].

Community of related nodes. With $\mathcal{S}_f$ and $\mathcal{T}_f$ denoting feature sets of structurally and textually related nodes, $\mathcal{M}_{S_f}^{n\times n}$ and $\mathcal{M}_{T_f}^{m\times m}$ define an adjacency matrix of the structural component and an affinity matrix based on the textual similarity component, respectively. Therefore, for each matrix there exist community sets ($\mathcal{K} \in \mathbb{R}^{n\times k}$, $\mathcal{Q} \in \mathbb{R}^{m\times q}$), such that $\{k_1, \cdots, k_n\} \in \mathcal{K}$ denotes possible communities in $\mathcal{M}_{S_f}$ and $\{q_1, \cdots, q_m\} \in \mathcal{Q}$ denotes communities in $\mathcal{M}_{T_f}$. For a matrix of reciprocal relationships $R \subseteq \mathcal{V} \times \mathcal{V}$[9] and the associated network data $\mathcal{D}$ ($\mathcal{V}, R \in \mathcal{D}$), there exist numerous communities $\{c_1, \cdots, c_k\} \in \mathcal{C}$ such that $\emptyset \subset c_i \subseteq \mathcal{V}$ and $\mathcal{C}$ denotes a community set. With any pair of similar nodes denoted by $v_i \sim v_j \iff \exists c_i \in \mathcal{C} : v_i, v_j \in c_i$, a more socially cohesive node community is formed by identifying overlapping nodes in both $\mathcal{K}$ and $\mathcal{Q}$, through a repetitive partition optimisation. Accordingly, the MCT framework contributes into

[8] For all experiments, $\tau \geq 0.5$, i.e. pairs are considered similar (1) if $\tau \geq 0.5$, otherwise dissimilar (0).
[9] R applies to both $\mathcal{M}_{S_f}$ and $\mathcal{M}_{T_f}$.

Algorithm 3 *: Algorithm MCT* identifies local communities known as *microcosms* in a network.

Initialisation: $\{\} \longleftarrow \mathcal{S}_r, \{\} \longleftarrow \mathcal{S}_u, \{\} \longleftarrow \mathcal{T}_r, \{\} \longleftarrow \mathcal{T}_u$
Input: a collection of network data $\mathcal{D}$
structural-component: ▷ *invoke f-sim (alg. 1)*
    f-sim$(\mathcal{D}) \mapsto \{\mathcal{S}_r, \mathcal{S}_u\}, \mathcal{M}_{A_{i,j}}^{n \times n}$ ▷ *alg. 1 output*
    textual-component: ▷ *invoke text-sim (alg. 2)*
        $\forall v_i \in \mathcal{S}_r$ get k tweets ▷ *set of texts* $\mathcal{T}_{v_i}$
        text-sim$(\mathcal{S}_r) \mapsto \{\mathcal{T}_r, \mathcal{T}_u\}, \mathcal{M}_{ta}^{m \times m}$ ▷ *alg. 2 output*
        compare all topics$(\mathcal{T}_{v_i} \in \mathcal{S}_r)$ using Eq. 11 ▷ *affinity matrix*
local clusters:
    $\psi(\mathcal{S}_r, \mathcal{T}_r)$ ▷ $\mathcal{S}_r, \mathcal{T}_r \geq \tau$
Output:
    $\mathcal{C}_{c_i}^{m \times p}$ ▷ *local communities*

two operational categories: (1) *optimising matrices of values* and (2) *optimising intra-cluster similarity*. The MCT can be considered as a multivariate function, made up by structural and textual components, allowing to define an objective function that maximises the overall joint similarity.

3.3.1. Optimising matrices of values

Recall that the set of textually related nodes $\mathcal{T}_r$ is a subset of the structurally related nodes $\mathcal{S}_r$ ($\mathcal{M}_{c_{vr}}$), i.e $\mathcal{T}_r \subseteq \mathcal{S}_r$. Since the optimisation goal is to *maximise* $\mathcal{T}_r$ ($\mathcal{M}_{vt}$), the two are equated under the constraint: $\mathcal{M}_{vt} = \mathcal{M}_{c_{vr}}$, such that $\mathcal{M}_{vt} - \mathcal{M}_{c_{vr}} = 0$. Noting the constraint, the simplified representation used in Eq. 9 also applies to $\mathcal{M}_{vt}$, given by $\mathcal{M}_{vt} = \mathcal{M}_{c_{vr}} = P$, to achieve the maximum values possible by determining the extremum of the function. Thus, the goal is to maximise the joint models under the constrained function according to Eq. 13[10]:

$$\psi_{st} = ||D - PQ^T||_F^2 + \sum_{i=1}^{n}\sum_{s=1}^{k} p_{is}\alpha_{is} + \sum_{j=1}^{p}\sum_{s=1}^{k} q_{js}\beta_{js} - \lambda Tr(\mathcal{M}_{vt}^T \mathcal{M}_{va} \mathcal{M}_{vt}) \tag{13}$$

3.3.2. Optimising intra-cluster similarity

Intra-cluster similarity optimisation is similar to the approach in Section 3.3.1 through the use of value matrices, but different objective function. The approach in Eq. 9 and the corresponding update rule (Eq. 10) are based on a matrix factorisation, which poses challenges with respect to exact or one-one mapping to the textually related clusters ($\mathcal{T}_r$). We know that the two are related at a higher level, since $\mathcal{T}_r \subset \mathcal{S}_r$, but the details about the shared clusters are not fully established. To address this challenge, we propose the following approach based on the node similarity. Information about similar nodes is stored in the nodes' affinity matrix ($\mathcal{M}_{va}^{n \times n}$), in which the magnitude of pairwise similarity decides entries in the matrix. Nodes are assigned to communities based on their degree of similarity denoted by $\mathcal{M}_{C_{vr}}^{n \times k}$ (of n nodes and k reciprocal-communities). For example, $\{c_{vr_1}, \cdots, c_{vr_k}\} \in \mathcal{C}_{vr}$ represents a set of nodes-reciprocal communities and membership in a cluster is qualified by Eq. 1 - 3. With a higher probability of forming a tie for nodes in the same cluster, community detection is based on optimising the joint

[10]The proportionality constant λ in Eq. 13 denotes a Lagrange multiplier.

Algorithm 4 *: Algorithm MCT-2* identifies local communities known as *microcosms* in a network.

Input: a collection of network data $\mathcal{D}$
structural-component:
    f-sim($\mathcal{D}$) $\mapsto \{\mathcal{S}_r, \mathcal{S}_u\}, \mathcal{M}_{sa}^{n\times n}$
    textual-component:
        $\forall v_i \in \mathcal{S}_r$ get k tweets
        text-sim($\mathcal{S}_r$) $\mapsto \{\mathcal{T}_r, \mathcal{T}_u\}, \mathcal{M}_{ta}^{n\times n}$
        compare all topics($\mathcal{T}_{v_i} \in \mathcal{S}_r$) using Eq. 11
Clusters initialisation:
    select four random seed nodes: $v_i, v_j, v_k, v_l \in \mathcal{S}_r$ and $v_i, v_j, v_k, v_l \in \mathcal{T}_r$
    compute pairwise similarities among v_i, v_j, v_k, v_l using $\psi_{s_r t_r}(v_i, v_j)$
if $\mathcal{T}_{sim}(\mathcal{T}_{v_i}, \mathcal{T}_{v_j}) \geq \tau$ **then**
    create single cluster $\mathcal{C}_{ij}$
else
    create two clusters $\mathcal{C}_i, \mathcal{C}_j$
end if
repeat 9 – 15 until$|\mathcal{C}_{ik}|_{k=1}^{M} = M$ ▷*maximum clusters M*
Assign nodes to clusters:
    $\forall v_i \in \mathcal{S}_r$ compute similarity with cluster's mean
    $max_{\phi(v_i, \mu_{C_i})}$ ▷ *assign* v_i *to the most similar* μ_{C_i}
update cluster's mean: $\mu_{C_i} \longleftarrow \mu_{C_i}$
Output:
    local communities

similarities of $\mathcal{S}_r$ and $\mathcal{T}_r$:

$$\psi_{st}(v_i, v_j) = (\lambda) \cdot \mathcal{S}_r(v_i, v_j) + (1 - \lambda) \cdot \mathcal{T}_r(v_i, v_j) \tag{14}$$

The goal of Eq. 14 is to maximise the joint similarity between $\mathcal{S}_r$ and $\mathcal{T}_r$ according to an aggregation criterion inspired by [70], based on the similarity scores between pairs and a user-defined balancing parameter[11] λ, with values in $(0, 1)$. We follow the approach in [71] to find optimum value for the λ. Algorithm 4 describes how nodes are assigned to relevant clusters until the stopping criterion, a user-defined integer M signifying the desired number of clusters, is reached.

4. Experimentation

This section presents our experimentation to evaluate the MCT against other existing methods.

4.1. Datasets

We utilise the following diverse datasets for the experimentation.

4.1.1. Ground-truth and predicted data

Unlike previous studies in which datasets from various social networks were collected [31, 72, 73], this study focuses on nodes with reciprocal, not directed, ties. The reciprocal

[11] Note that this is different from the one used in optimisation based on matrices of values.

Algorithm 5 *Algorithm search-dyads* profiles users with directed and undirected ties on Twitter

Initialisation: $1 - edge \longrightarrow \{\}, dyads \longrightarrow \{\}$
Input: begin with an arbitrary set of seed users, say k
while $k \neq \emptyset$ **do**
    $\forall v_i \in k$, get sets of *friends* fr_{v_i}, *followers* fl_{v_i}, $fr_{v_i}, fl_{v_i} \in m_{v_i}$; m_{v_i} denotes v_i network
    $\forall v_j \in fr_{v_i}$, retrieve the sets fr_{v_j} and fl_{v_j}, $fr_{v_j}, fl_{v_j} \in m'_{v_j}$; m'_{v_j} denotes v_j network
    if $v_i \in fr_{v_j}$ **then**
        $v_i \sim v_j$ ▷ both follows one another
        update dyads
    else
        v_i follows v_j
        update 1-edge
    end if
end while

collection consists of *dyadic* and *transitive* datasets, which were collected using Twitter's Application Programming Interface (API) according to Algorithm 5. The process returns a collection of *tweet objects*, a complex object with many descriptive fields, which allows to extract structural and textual components for analysis. The collection begins with a search on the network profile of each from a finite set of seed users[12], or a network composition m_{v_i}, consisting of lists of friends fr_{v_i} and followers fl_{v_i}, to determine user pairs that follow each other. The set of reciprocal pairs is denoted by $\kappa \in m_{v_i}$ and the transitive dataset is a scaled-version of dyadic data.

In addition to the collection of nodes with actual pairwise ties (denoted as G-pTie in Table 4.1), the ground-truth dataset also consists of public data associated with COVID-19 outbreak (G-pMention) related to aspects of scepticism and myths about the pandemic [75]. The data contains two broad categories: information put forward by credible sources, such as the World Health Organisation (WHO), and information from users dismissing WHO's guidelines on combating the pandemic. The dataset consists of interaction information about users who mention each other. Nodes with frequent mentioning are highly likely to be in the same community. For the dataset consisting of predicted pairwise ties (P-pTie), a reciprocal tie exists between v_i and v_j if $p(R_{v_i,v_j}) \geq \tau$, otherwise just a directed tie. In Table 2, SND1 refers to synthetic network data generated based on LFR approach (see Section 4.3.2 for details).

4.1.2. Public datasets

To reinforce evaluation and generalisation, we use the following collections of publicly available datasets. The datasets consist of real-world networks commonly used for community detection. Essentially, the following datasets have been used: Zachary's karate club [76], dolphin social network [77], political blog dataset [78] and Ego-network, consisting Facebook and Twitter datasets [79]. The Facebook data contains anonymised *'circles'* or *'friends lists'*, and *node features (profiles)*. Each node has *node ids*, sets of *connections* or *edges*, and *anonymised features* encoding information about its *circle*. The Facebook data allows to explore communities using each user's network circle in terms of size and diversity of membership. Moreover, the

[12] Seed users are verified or unverified accounts devoid of spammers or social bots were collected by SPD filtering [74]. A *'list'* on Twitter allows a user to store a set of preferred users and can be used to obtain relevant information.

collection consists of *synthetic data*, which is based on the approach proposed in [71] to generate synthetic network with known parameters. The synthetic nature of the networks makes it possible to explore the parameters' space for the best community structure in the network. Table 2 shows basic statistics of datasets used in this study.

4.2. Meta-analysis

Owing to the prevalence of unreciprocated and event-type ties on Twitter, we conjecture that mining tasks, such as community detection, are less effective and more challenging. In this section, our goal is to apply a pragmatic approach that provides a statistical analysis of relevant metrics in the datasets to identify strongly correlated node attributes (Figure 4(b)) with reciprocity among nodes. The empirical cumulative distribution function (ECDF) gives the probability of a quantity evaluated at arbitrary points. We use it to analyse observations, such as the variation of dyads or Simmelian ties across user categories or network size.

Table 2: A summary of microcosms detection datasets. V and E denote the node and edge size, respectively. G-pTie and P-pTie denote groundtruth and predicted sets of users with pairwise connectivity; G-Mention denotes collection of users with pairwise mentioning; $\mu_{deg.}$ refers to the average degree in each data category.

	Category	V	E	Description
Datasets	G-pTie	18973	15538	$\mu_{deg.} = 1.6379$
	P-pTie	15038	1298998	$\mu_{deg.} = 172.7621$
	G-pMention	514	259	$\mu_{deg.} = 1.0078$
	Karate club	34	78	Consists of 2 groundtruth communities
	Dolphin	62	—	Consists of 2 groundtruth communities
	Pol. Blog	1224	—	Consists of 2 groundtruth communities
	ego-Facebook	4039	88234	$\mu_{deg.} = 43.6910$
	ego-Twitter	81,306	1768149	—
	SND1	1000	—	—

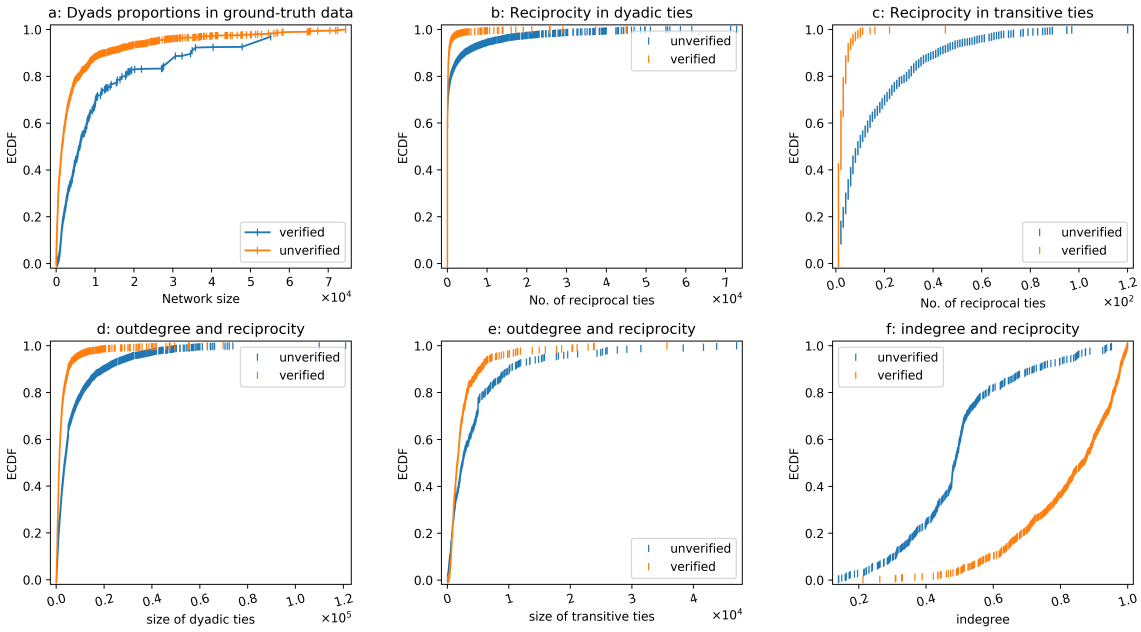

Figure 6: Sub-figures (a), (b) and (c) show the reciprocity effect on nodes with many dyadic relationships. Sub-figures (d), (e) and (f) show outdegree-reciprocity and indegree-reciprocity relationship in the ground-truth data.

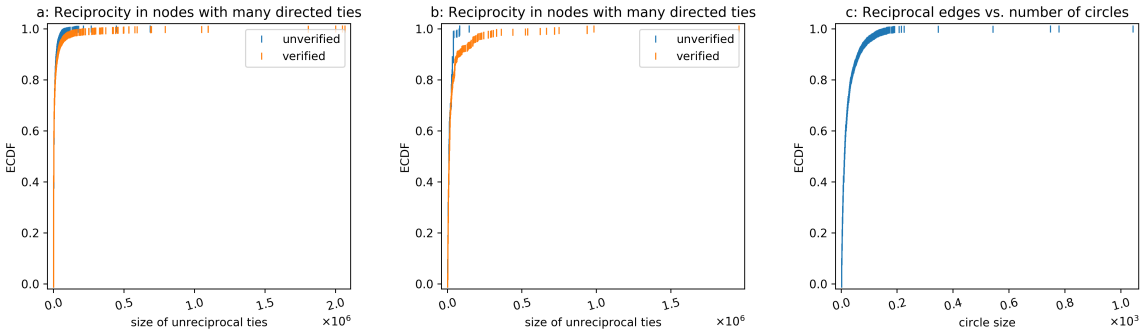

Figure 7: Relationship between the number of directed ties and reciprocity across different nodes in the datasets.

4.2.1. Proportion of reciprocal units

Noting the flexibility of connections and the rarity of reciprocal links on Twitter, large scale dyadic ties are rare and difficult to locate. Using Algorithm 5, we collected *directed* or *1-edge* and *undirected* data, and examined the *network topology* of each category and its utility in the detection of local communities. In Figure 6, there is a high proportion of reciprocity in unverified users in comparison to the verified counterpart. The reciprocity ability slightly decreases with increasing network size of the user, which can be attributed to the difficulty in keeping track of and responding to all followership requests. Sub-figures 6(a) and (b) show the relationship between reciprocity and the number of reciprocal ties. While there is higher reciprocity in the unverified category, the verified category shows almost 100% reciprocity with a relatively small network size. Sub-figure 6(c) shows similar behaviour in transitive ties, but is more evident in the unverified users category. Similarly, sub-figures 6(d),(e) and (f) show the relationship between outdegree and reciprocity and indegree and reciprocity in the ground-truth data. The behaviour resembles an inverse relationship in which reciprocity decreases with increasing outdegree (sub-figures 6(d) and (e)). Sub-figure 6(f) shows an almost linear relationship between indegree and reciprocity, especially among the unverified category. In the verified category, the effect is low and seems to shoot once the network size increases (vis-a-vis indegree/number of followers). There is an instant reciprocity in the unverified users, which can be explained by suggesting that the users are interested in expanding their network. Figure 7 shows the relationship between the number of directed ties and reciprocity across user categories in the data. The results demonstrate that verified users have many directed nodes or unreciprocated ties but with less reciprocity. This observation holds for nodes with many dyadic and transitive ties in the data.

4.3. Evaluation

To ascertain the efficacy and relevance of the study's output, evaluation entails thorough analyses and comparison with relevant baselines drawn from the literature. Quantitative analyses of experimentation on various datasets using the baseline algorithms is involved. Other forms of evaluation are specific to the structural and textual levels of the MCT strategy. The evaluation process aims to: (a) investigate the effect of utilising structurally-related nodes in identifying local communities in social networks, (b) compare structurally-related clusters with textually-related clusters, and (c) evaluate the performance of MCT in comparison with baseline models.

4.3.1. Evaluation metrics

This section discusses quantitative measures for validating the performance of MCT and baseline methods. Because of the multilevel approach, the metrics are suitable for evaluating network structure (structural clusters) and textual-clusters (roughly considered as labels).

Clustering coefficient and Community cohesion. Clustering coefficient, C_{coeff}), is used to quantify the clustering tendency of a given node in relation to other nodes within a network [38]. Computing C_{coeff} requires: $edges = \frac{k_i(k_i-1)}{2}$ and $C_{coeff_i} = \frac{2E_i}{k_i(k_i-1)}$ where i, k_i, E_i denote a network node, the number of edges connecting i to k_i other nodes in the network, and the actual number of existing edges between k_i nodes, respectively. The ratio $E_i \propto \frac{k_i(k_i-1)}{2}$ defines the clustering coefficient of a node. *Community cohesion* demonstrates the level of connectivity within a community and is captured by measuring the degree of cohesiveness. Due to the presence of a strong connectivity among nodes, a well-connected community is intuitively difficult to divide into sub-communities [80]. Any useful metric that reveals the degree of cohesion can be used to evaluate cohesiveness, i.e., if the community is well-connected and difficult to partition. In this study, cohesiveness is measured by the degree of similarities $\mathcal{S}_r$ and $\mathcal{T}_r$. We compute the *average degree* (μ_{deg}), defined as the average node degree to other member nodes [81]. Moreover, we use the *accuracy metric*, i.e., the fraction of predicted labels to the total number of data points.

Modularity and NMI. Modularity, $\mathbf{Q}$, measures the strength of communities, as the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random [42]. Usually, $\mathbf{Q} > 0$ signifies the possible presence of a community structure and the higher the values the better [41]. Normalised Mutual Information (NMI) is another statistical tool to evaluate the quality of network clusters [82]. NMI measures the degree of agreement between network partitions, based on the assumption that each node in a community, $v_i \in \mathcal{V}$, is associated with both the *true community* and the *predicted community*, such that $l_{v,p} = i$ defines the predicted community i of a node [83]. Furthermore, we apply *Rand* and *Jaccard* similarity metrics, which are based on tracking both correctly and incorrectly classified pairs of nodes, especially in groundtruth datasets.

4.3.2. Baseline Models

For evaluation, *MCT* is applied alongside the following detection algorithms with different modes of operation on the datasets described in Section 4.1 to identify local community structures.

Girvan-Neuman (G-N) and Label propagation (LP). The G-N algorithm assumes that a community detection algorithm can naturally detect divisions among vertices without external influence or imposed restrictions on the divisions [5]. Accordingly, Girvan and Newman [84] proposed the iterative G-N algorithm that progressively removes network edges based on betweenness, a metric to quantify traffic flow among nodes. Each node's betweenness score dictates which edge to remove. The most critical nodes are likely to experience high traffic flow, hence will possibly create a bottleneck. The LP algorithm is an iterative clustering method that converts unlabelled data to labelled given an initial seed of labelled data. Labelling involves a repetitive random node reshuffling and tagging with the most frequent label among its neighbours until convergence [85]. The labelled data information is then propagated across the whole network.

Synthetic Network Model. This is achieved using the widely used approach, or LFR model, proposed in [86] to generate synthetic networks with planted partitions or community structures. For a given network G generated via the LFR, the following basic model's parameters are defined: $\gamma, \beta, \bar{d}, \hat{\mu}$ denoting exponents of the power-law degree distribution, community size distribution, mean degree and mixing parameter, respectively. Accordingly, the model ensures that nodes' degrees are sampled independently whose distribution exhibit power-law behaviour and the mixing parameter, $\hat{\mu}$, to distribute nodes' indegree and outdegree such that $1 - \hat{\mu}$ and $\hat{\mu}$ denote the proportions of edges shared with nodes in the same and different communities, respectively. The SND1 netwrok in Table 2 is generated based on the LFR approach. The network consists of 1000 nodes, $\gamma = 1.5$, $\bar{d} = 15$, $\gamma_C = 0.8$, $C_{min} = 30$, $C_{max} = 300$, and the mixing parameter ,$\hat{\mu}$, sampled from $0.1, 0.01, 0.2, 0.3, 0.5, 0.7, 0.9, 1$. Because the parameters pertaining the network and the embedded community structure are known, relevant community detection methods should be able to detect or identify values (especially for the community) that approximate such parameters.

The Planted Partition Model (PPM) is a form of likelihood optimisation algorithms that are commonly used for community detection task. Due to their mathematical efficacy, many algorithms are defined based on relevant assumptions about the underlaying structure in the network. Under this approach, a network is a composition of communities, which are used to infer the network [87]. The PPM relies on the community membership of nodes to probabilistically decide whether any pairs of nodes are connected. We apply variant of the PPM (degree-corrected planted partition model [88]) and an extended version of the LFR model proposed in [71] as part of the evaluation.

4.4. Detection of community structure

In this section, we focus on the detection of community structures using our proposed method[13], introduced in Section 3.3, and the baseline models, described in Section 4.3.2. The detection process consist of four steps: (1) retrieve a set of nodes with reciprocal ties on Twitter, (2) compute the similarity proportion between pairs, using Algorithm 1, (3) compare prediction accuracy using the ground-truth, and (4) perform clustering for community detection.

4.4.1. Effectiveness of tie prediction

Using Algorithm 1, which computes the similarity between the corresponding features of any pairs of nodes, we report its efficacy in the prediction pipeline. Due to the availability of empirical data, the effectiveness of the model is quantified with respect to the degree of conformity with the ground-truth data. This is vital because the tie prediction segment is not relevant if it does not add value to the overall detection framework. The accuracy of the prediction is obtained by computing the ratio of predicted reciprocal ties to true reciprocal ties. The best result achieved is $.608$ accuracy; depending on the threshold τ, the accuracy may be lower or higher. Sub-figure 8(c) shows possible values of τ and the corresponding accuracy.

4.4.2. Community structure

We examine how the use of a collection of structurally-related nodes affects community detection, and compare performance. For the experiments, we apply the proposed method,

[13] relevant datasets and implementation code available at https://github.com/ijdutse/mct

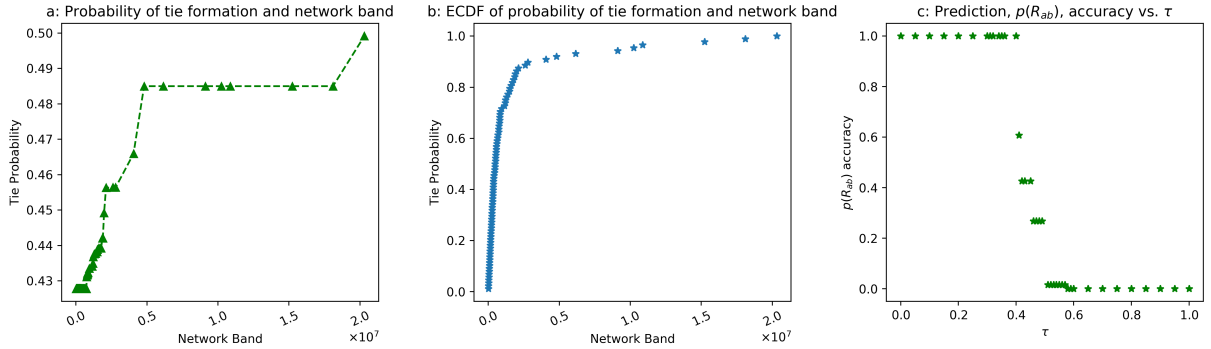

Figure 8: Sub-figures a and b show the probability of tie formation as a function of *network size*; there is a high chance of reciprocating a tie among users in a network band of 0.5×10^7. Sub-figure c shows the prediction accuracy versus the threshold value in *f-sim* (Algorithm 1). The prediction accuracy is almost 100% when the value of the threshold is low; conversely, the accuracy is almost 0% when the threshold value is very high. A *switch-point* can be observed toward the midpoint in which the accuracy is just above 60% at a threshold value of about 0.41. With additional features, the prediction can be improved. For instance, the inclusion of a description feature led to a significant improvement; however, it requires training on a large corpus to obtain the embedding of terms in text.

(MCT), and the baselines, G-N [84] and LP [85]. Table 3 shows the results of applying the community detection algorithms on the data according to the evaluation metrics described in Section 4.3.1. Although all the algorithms detected community structures, there are quantitative variations among the outcomes. Our analysis is along the following dimensions.

Effect of datasets: All algorithms perform best on the ground-truth data, followed by *ego-Facebook*, then *predicted*, and worst on *ego-Twitter*. The *ego-Facebook* data consists of nodes with reciprocal ties, but the textual feature set is small, making it less complex than the other datasets. We consider *homophily* and *structural equivalence* as precursors of communities, in which nodes with similar profiles or social status are more likely to interact and establish a small community. For instance, sub-figures 8(a) and (b) show homophily as a form of structural equivalence based on network size and indegree for examining the probability the formation of an edge. Sub-figures 8(a) and (b), depicts a behaviour that resembles an inverse relationship: increase in network size results in decrease in reciprocity.

Effect of models: Table 3 also demonstrates the performance of each model. The *MCT* results indicate a more localised structure noting the magnitude of *Q*, *NMI* and the number of detected communities (*#DC*) with respect to the *ground-truth* data. We attribute the improvement to the use of in-depth structural features that introduce a connectivity layer. *MCT* explores the data for community structures at local and global level through a high-level grouping of nodes into

Table 3: Results of experiments on three datasets for community detection using algorithms based on structural properties. G-N: Girvan–Neuman, LP: Label Propagation, MCT: Multilevel Clustering Technique, #DC: Number of Detected Communities

Dataset	G–N			LP			MCT		
	Metric		#DC	Metric		#DC	Metric		#DC
	Q	NMI		Q	NMI		Q	NMI	
G-pTie	.908	.794	308	.77	.602	1319	.915	.791	263
ego-Twitter	.334	.197	1431	.215	.131	2131	.307	.230	1131
ego-Facebook	.522	.590	1037	.421	.304	1780	.503	.372	1845
P-pTie	.473	.311	1107	.360	.267	2071	.601	.472	985

communities, according to the network size and the recognition of bi-modal information sources.

For the parameterised approach, groundtruth datasets form the basis of the evaluation. Therefore, we discuss the results obtained in Table 4 using Rand and Jaccard scores as the evaluation metrics. Generally, the results in the Table indicate good performance, especially with respect to the Rand score and on small datasets such as the Karate club. The values associated with the MCT signify better performance across all the datasets. However, there are instances where the algorithm's performance lags behind. For instance, the PPM and ILFR perform better on the Karate and Pol. Blogs datasets, which we attribute to the small size nature of the datasets. Moreover, there is a significant improvement on performance on the synthetic datasets, i.e. SND1. This s expected since the network consists of well defined community structure. A common trait among the algorithms is that they perform poorly based on the Jaccard index, suggesting that the metric is somewhat strict or further optimisation is needed.

5. Discussion

In this section, we discuss some significant observations from the study.

Impact of reciprocal units and text aggregation for clustering. One of the assumptions of this is that recognising a set of reciprocal units for community detection offers a more cohesive community representation. Since small groups allow modular analysis of social networks [89, 90], we examined reciprocal ties, dyadic and Simmelian, as the basic units of relational interaction on Twitter. However, Twitter's flexible and eccentric connections entangle locating nodes with reciprocal links. Structural similarity allows to organise nodes into connected clusters and simplifying community detection. Structurally similar nodes are more likely to connect and belong to the same community. The high volume and small size of tweets make comparisons of discussions context challenging. Because a single tweet may not yield enough information about a discussion, we need to balance between quantity and quality. We collected a finite set of tweets from each node v_i that defines a user corpus $\mathcal{T}_{v_i}$, and computed its overall theme to compare with other nodes. Textually-related nodes $\mathcal{T}_r$ are identified by a topic modelling technique that compares the similarities of the discussion topics of structurally similar nodes.

Improving social cohesion in the detection task. Online content increases rapidly in volume and complexity and is dominated by influential users. These facts make the detection of socially cohesive groups on Twitter challenging. With respect to *sociometry*, the formation of a social tie can be based on *event-type* or *state-type ties*. The size of a network and the size of communities

Table 4: Results of experiments on three datasets for community detection using relevant algorithms and evaluation metrics.

	Model	PPM		LFR		MCT		ILFR	
	Metric	Rand	J	Rand	J	Rand	J	Rand	J
Datasets	G-pMention	0.501	0.005	0.501	0.005	0.662	0.169	0.501	0.005
	Karate club	0.734	0.475	0.734	0.475	0.701	0.461	0.720	0.425
	Dolphin	0.651	0.379	0.581	0.255	0.696	0.340	0.536	0.167
	Pol. Blob	0.903	0.821	0.878	0.775	0.565	0.487	0.581	0.166
	G-pTie	0.621	0.012	0.621	0.012	0.629	0.210	0.621	0.012
	P-pTie	0.652	0.001	0.652	0.001	0.696	0.237	0.652	0.001
	SND1	0.846	0.431	0.795	0.395	0.879	0.510	0.601	0.317
	ego-Facebook	0.683	0.317	0.597	0.301	0.697	0.332	0.579	0.298

are almost linearly correlated. Similarly, the size of a network is inversely correlated with its degree of homogeneity. The degree of interaction is higher among structurally similar users. Often, users that discuss with and mention each other are engaged in reciprocal ties, showing strong social cohesion. Based on the idea of social *homophily*, users with many reciprocated ties are crucial in analysing socially cohesive groups. Figure 1 shows that communities on Twitter can be formed in many ways. A bi-modality approach differs across networks with respect to the depth of the features associated with the structural and textual modalities [29, 31, 16]. Bi-modalities, e.g., network structure, features and attributes of nodes, lead to better and more interpretable community detection results. In Twitter, the structural component is not fully captured as it relies on directed connections. *MCT* exploits the usability of features in the detection of a local community through the impact analyses of both modalities, especially the structural one. We have shown that a structural component is useful in community detection and has minimal practical requirements. *MCT* offers a compact way to find and represent co-occurring users or user groups, allowing to explore local and global clustering requirements.

6. Conclusion

Many natural networks exhibit a certain degree of organisation, in which node groups form tightly connected units called communities. Community detection allows to understand the network structure and extract useful information. Detecting socially cohesive communities on Twitter is still challenging. While many methods have been proposed, they often discover disparate communities, likely to be socially unrelated. We observed that the topology of eccentric connections contributes to the detection of socially unrelated users and encourages the propagation of spurious content. Consequently, we propose a *multilevel clustering technique (MCT)* to identify socially cohesive user groups, i.e. *microcosms*, on Twitter.

The proposed *MCT* framework, jointly modelling structural and intrinsic textual features, contributes toward a methodological paradigm for cohesive community detection in a dynamic and heterogeneous social media. This is important because until recently, community detection algorithms focused on single modality, e.g. using node attributes or connectivity. Recent studies that combine information modalities are limited in capturing the nuances and intricate connection structure in platforms, such as Twitter. To improve the identification of socially cohesive communities, *MCT* offers a scalable detection strategy. The approach addresses the problem of structurally unrelated users, by adding a layer of social cohesion to existing community detection methods. In summary, *MCT* contributes: (1) a systematic exposition of community detection or clustering algorithms, (2) an in-depth utilisation of the bi-modality for community detection, and (3) detection of network communities at various levels.

A note on the proposed method's complexity is in order here. When the network size is huge, it is challenging to authoritatively specify when a given community detection algorithm will converge. Thus, we rely on a single iteration to analyse the algorithm's complexity, which will provide insights to its future performance. Let us assume that the execution complexity of a basic parameterised algorithm is $f(C)$, then the term $O(f(C) \times s \times r \times m)$, where *s* is the number of comparisons in deciding the next cluster, *r* is the size size and *m* the number runs. Execution wise, the complexity of the algorithm is relatively low. However, it tends to increase with growing data-points, hence the needs for further improvement in future work.

Acknowledgements

The third author has participated in this research work as part of the *TYPHON* Project, which has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 780251.

Appendix A: Supplementary information

Structural Communities: optimisation and interpretability. Recall that the *network-communities* ($\mathcal{M}_{C_{ns}}^{n\times p}$) matrix is decomposed into its approximate constituents given by Eq. 7, i.e. $\mathcal{M}_{C_{ns}} \approx \mathcal{M}_{vr}\mathcal{M}_{C_{nr}}^{T}$ and the optimisation function (Eq. 8) given by: $min._{\mathcal{M}_{c_{vr}},\mathcal{M}_{c_{nr}}} ||\mathcal{M}_{c_{ns}} - \mathcal{M}_{c_{vr}}\mathcal{M}_{c_{nr}}^{T}||_F^2$ subject to $\mathcal{M}_{c_{vr}}, \mathcal{M}_{c_{nr}} \geq 0$. The following conventions are used to represent the matrices: $\mathcal{M}_{c_{ns}} \mapsto D$, $\mathcal{M}_{c_{vr}} \mapsto P = [p_{is}]$, $\mathcal{M}_{c_{nr}} \mapsto Q = [q_{js}]$. We follow the *NMF* scheme [64] in the *modelling of structural communities*.

Iterative Update. In response to the additional parameters (α, β with values ≤ 0,) induced by the *Lagrangian relaxation*, the objective function M_{sr} is given by the following equation:

$$M_{sr} = ||D - PQ^T||_F^2 + \sum_{i=1}^{n}\sum_{s=1}^{k} p_{is}\alpha_{is} + \sum_{j=1}^{d}\sum_{s=1}^{k} q_{js}\beta_{js}$$

To solve the optimisation problem, the process begins with computing the gradient of the Lagrangian relaxation with respect to the first aspect of the *minmax* (i.e. minimisation) optimisation variables. To achieve an optimal solution, the optimisation condition needs to be based on P, Q only. Hence, to eliminate the introduced *Lagrangian multipliers*, the *KKT optimality condition*, which suggests that $p_{is}\alpha_{is} = 0$ and $q_{js}\beta_{js} = 0$, is applied. We then solve for the optimisation parameters as follows.

$$\begin{aligned} ||D - PQ^T||_F^2 &= (D - PQ^T)^T(D - PQ) \\ &= (D^T - P^TQ)(D - PQ) \\ &= \underbrace{D^TD}_{1} - \underbrace{D^TPQ}_{2} - \underbrace{QP^TD}_{3} + \underbrace{QP^TPQ}_{4} \end{aligned} \tag{15}$$

In Eq. 15, the second term (2) and third term (3) are equal and the fourth term (4) can be expressed in a quadratic form depending on the parameter of interest (for minimisation). Thus,

$$M_{sr} = D^TD - 2QP^TD + P^2Q^TQ + \sum_{i=1}^{n}\sum_{s=1}^{k} p_{is}\alpha_{is} + \sum_{j=1}^{d}\sum_{s=1}^{k} q_{js}\beta_{js} \tag{16}$$

From Eq. 16, the partial differentiation with respect to P gives:

$$\frac{\partial}{\partial p_{is}} M_{sr} = -(2DQ)_{is} + (2PQ^TQ)_{is} + \alpha_{is}$$

divide by 2 and equate to zero

$$= -(DQ)_{is} + (PQ^TQ)_{is} + \alpha_{is} = 0$$

to eliminate the relaxation parameters multiply with p_{is} throughout

$$= -(DQ)_{is}p_{is} + (PQ^TQ)_{is}p_{is} + \alpha_{is}p_{is} = 0 \tag{17}$$

the term $\alpha_{is}p_{is}$ equates to 0 according to KKT optimality, thus

$(PQ^TQ)_{is}p_{is} = (DQ)_{is}p_{is}$

the update rule:

$$p_{is} = \frac{(DQ)_{is}p_{is}}{(PQ^TQ)_{is}}$$

The last term or expression in Eq. 17 is the update rule for the parameter P. A similar process applies to the parameter Q:

$$M_{sr} = D^TD - 2QP^TD + P^TPQ^2 + \sum_{i=1}^{n}\sum_{s=1}^{k} p_{is}\alpha_{is} + \sum_{j=1}^{d}\sum_{s=1}^{k} q_{js}\beta_{js} \tag{18}$$

The partial derivative with respect to Q is given by the following:

$$\frac{\partial}{\partial q_{js}} M_{sr} = -(2D^TP)_{js} + (2QP^TP)_{js} + \beta_{js}$$

divide by 2 and equate to zero

$$= -(D^TP)_{js} + (QP^TP)_{js} + \beta_{js} = 0$$

to eliminate the relaxation parameters multiply with q_{js} throughout

$$= -(D^TP)_{js}q_{js} + (QP^TP)_{js}q_{js} + \beta_{js}q_{js} = 0 \tag{19}$$

the term $\beta_{js}q_{js}$ equates to 0 according to KKT optimality, thus

$(QP^TP)_{js}q_{js} = (D^TP)_{js}q_{js}$

the update rule:

$$q_{js} = \frac{(D^TP)_{js}q_{js}}{(QP^TP)_{js}}$$

The last term or expression in Eq. 19 is the update rule for the parameter Q. The process of updating P, Q involves comparing their values to the original matrix D, and the goal is to minimise the difference or *error*. The iterative update of the parameters (p_{is} and q_{js}) continues until convergence.

References

[1] A. Lancichinetti, S. Fortunato, J. Kertesz, Detecting the overlapping and hierarchical community structure in complex networks, New journal of physics 11 (2009) 033015.

[2] J. Scott, Social network analysis, Sociology 22 (1988) 109–127.

[3] D. J. Watts, S. H. Strogatz, Collective dynamics of 'small-world'networks, nature 393 (1998) 440.

[4] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Reviews of modern physics 74 (2002) 47.

[5] M. E. Newman, M. Girvan, Finding and evaluating community structure in networks, Physical review E 69 (2004) 026113.

[6] B. Berelson, G. A. Steiner, Human behavior: An inventory of scientific findings. (1964).

[7] M. E. Shaw, Group dynamics: The psychology of small group behavior (1971).

[8] M. Granovetter, Problems of explanation in economic sociology, Networks and organizations: Structure, form, and action (1992) 25–56.

[9] D. J. Brass, K. D. Butterfield, B. C. Skaggs, Relationships and unethical behavior: A social network perspective, Academy of management review 23 (1998) 14–31.

[10] M. E. Newman, J. Park, Why social networks are different from other types of networks, Physical review E 68 (2003) 036122.

[11] R. J. Williams, N. D. Martinez, Simple rules yield complex food webs, Nature 404 (2000) 180.

[12] G. W. Flake, S. Lawrence, C. L. Giles, F. M. Coetzee, Self-organization and identification of web communities, Computer (2002) 66–71.

[13] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, P. Spyridonos, Community detection in social media, Data Mining and Knowledge Discovery 24 (2012) 515–554.

[14] C. Cao, Q. Ni, Y. Zhai, An improved collaborative filtering recommendation algorithm based on community detection in social networks, in: Proceedings of the 2015 annual conference on genetic and evolutionary computation, ACM, 2015, pp. 1–8.

[15] M. E. Newman, Properties of highly clustered networks, Physical Review E 68 (2003) 026121.

[16] J. Yang, J. McAuley, J. Leskovec, Community detection in networks with node attributes, in: 2013 IEEE 13th International Conference on Data Mining, IEEE, 2013, pp. 1151–1156.

[17] V. Arnaboldi, A. Guazzini, A. Passarella, Egocentric online social networks: Analysis of key features and prediction of tie strength in facebook, Computer Communications 36 (2013) 1130–1144.

[18] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, et al., Global landscape of protein complexes in the yeast saccharomyces cerevisiae, Nature 440 (2006) 637.

[19] M. A. Nascimento, J. Sander, J. Pound, Analysis of sigmod's co-authorship graph, ACM Sigmod record 32 (2003) 8–10.

[20] G. Palla, A.-L. Barabási, T. Vicsek, Quantifying social group evolution, Nature 446 (2007) 664.

[21] E. O. Laumann, P. V. Marsden, D. Prensky, The boundary specification problem in network analysis, Research methods in social network analysis 61 (1989) 87.

[22] S. P. Borgatti, D. S. Halgin, On network theory, Organization science 22 (2011) 1168–1181.

[23] H. Kwak, C. Lee, H. Park, 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.

[24] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, B. Y. Zhao, User interactions in social networks and their implications, in: Proceedings of the 4th ACM European conference on Computer systems, Acm, 2009, pp. 205–218.

[25] J. Chen, O. R. Zaiane, R. Goebel, Detecting communities in large networks by iterative local expansion, in: 2009 International Conference on Computational Aspects of Social Networks, IEEE, 2009, pp. 105–112.

[26] P. Doreian, V. Batagelj, A. Ferligoj, Positional analyses of sociometric data, Models and methods in social network analysis 77 (2005) 77–96.

[27] J. Shi, J. Malik, Normalized cuts and image segmentation, Departmental Papers (CIS) (2000) 107.

[28] M. E. Newman, Modularity and community structure in networks, Proceedings of the national academy of sciences 103 (2006) 8577–8582.

[29] R. Balasubramanyan, W. W. Cohen, Block-lda: Jointly modeling entity-annotated text and entity-entity links, in: Proceedings of the 2011 SIAM International Conference on Data Mining, SIAM, 2011, pp. 450–461.

[30] W. Lin, X. Kong, P. S. Yu, Q. Wu, Y. Jia, C. Li, Community detection in incomplete information networks, in: Proceedings of the 21st international conference on World Wide Web, ACM, 2012, pp. 341–350.

[31] J. Leskovec, J. J. Mcauley, Learning to discover social circles in ego networks, in: Proceedings of NIPS, 2012, pp. 539–547.

[32] J. Allan, R. Papka, V. Lavrenko, On-line new event detection and tracking., in: Sigir, volume 98, Citeseer, 1998, pp. 37–45.

[33] Y. Yang, A study of thresholding strategies for text categorization, in: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2001, pp. 137–145.

[34] T. Brants, F. Chen, A. Farahat, A system for new event detection, in: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM, 2003, pp. 330–337.

[35] G. P. C. Fung, J. X. Yu, P. S. Yu, H. Lu, Parameter free bursty events detection in text streams, in: Proceedings of the 31st international conference on Very large data bases, VLDB Endowment, 2005, pp. 181–192.

[36] C. M. Bishop, Pattern recognition and machine learning, springer, 2006.

[37] P. Erdős, A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci 5 (1960) 17–60.

[38] D. J. Watts, P. S. Dodds, Influentials, networks, and public opinion formation, Journal of consumer research 34 (2007) 441–458.

[39] E. Katz, P. F. Lazarsfeld, E. Roper, Personal influence: The part played by people in the flow of mass communications, Routledge, 2017.

[40] H. Sundaram, Y.-R. Lin, M. De Choudhury, A. Kelliher, Understanding community dynamics in online social networks: a multidisciplinary review, IEEE Signal Processing Magazine 29 (2012) 33–40.

[41] M. E. Newman, Detecting community structure in networks, The European Physical Journal B 38 (2004) 321–330.

[42] M. E. Newman, Fast algorithm for detecting community structure in networks, Physical review E 69 (2004) 066133.

[43] D. J. Lawson, D. Falush, Population identification using genetic data, Annual review of genomics and human genetics 13 (2012) 337–361.

[44] C. D. Manning, C. D. Manning, H. Schütze, Foundations of statistical natural language processing, MIT press, 1999.

[45] C. Aggarwal, K. Subbian, Evolutionary network analysis: A survey, ACM Computing Surveys (CSUR) 47 (2014) 10.

[46] P. Pons, M. Latapy, Computing communities in large networks using random walks., J. Graph Algorithms Appl. 10 (2006) 191–218.

[47] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of statistical mechanics: theory and experiment 2008 (2008) P10008.

[48] A. Pothen, H. D. Simon, K.-P. Liou, Partitioning sparse matrices with eigenvectors of graphs, SIAM journal on matrix analysis and applications 11 (1990) 430–452.

[49] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (2003) 993–1022.

[50] X. Yan, J. Guo, Y. Lan, X. Cheng, A biterm topic model for short texts, in: Proceedings of the 22nd international conference on World Wide Web, ACM, 2013, pp. 1445–1456.

[51] P. Berkhin, A survey of clustering data mining techniques, in: Grouping multidimensional data, Springer, 2006, pp. 25–71.

[52] K. Chaudhuri, S. M. Kakade, K. Livescu, K. Sridharan, Multi-view clustering via canonical correlation analysis, in: Proceedings of the 26th annual international conference on machine learning, ACM, 2009, pp. 129–136.

[53] J. Liu, C. Wang, J. Gao, J. Han, Multi-view clustering via joint nonnegative matrix factorization, in: Proceedings of the 2013 SIAM International Conference on Data Mining, SIAM, 2013, pp. 252–260.

[54] S. Bickel, T. Scheffer, Multi-view clustering., in: ICDM, volume 4, 2004, pp. 19–26.

[55] G. Chao, S. Sun, J. Bi, A survey on multi-view clustering, arXiv preprint arXiv:1712.06246 (2017).

[56] M. Ester, R. Ge, B. J. Gao, Z. Hu, B. Ben-Moshe, Joint cluster analysis of attribute data and relationship data: the connected k-center problem, in: Proceedings of the 2006 SIAM International Conference on Data Mining, SIAM, 2006, pp. 246–257.

[57] Y. Zhou, H. Cheng, J. X. Yu, Graph clustering based on structural/attribute similarities, Proceedings of the VLDB Endowment 2 (2009) 718–729.

[58] M. S. Granovetter, The strength of weak ties, in: Social networks, Elsevier, 1977, pp. 347–367.

[59] I. Inuwa-Dutse, M. Liptrott, Y. Korkontzelos, Analysis and prediction of dyads in twitter, in: International Conference on Applications of Natural Language to Information Systems, Springer, 2019, pp. 303–311.

[60] A. Marley, M. Regenwetter, Choice, preference, and utility: Probabilistic and deterministic representations, New handbook of mathematical psychology 1 (2016) 374–453.

[61] I. Inuwa-Dutse, M. Liptrott, I. Korkontzelos, Simmelian ties on twitter: empirical analysis and prediction, in: 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), IEEE, 2019, pp. xx–xx.

[62] S. Wasserman, K. Faust, Social network analysis: Methods and applications, volume 8, Cambridge university press, 1994.

[63] J. Han, J. Pei, M. Kamber, Data mining: concepts and techniques, Elsevier, 2011.

[64] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788.

[65] C. C. Aggarwal, Machine learning for text, Springer, 2018.

[66] D. P. Bertsekas, Nonlinear programming, Journal of the Operational Research Society 48 (1997) 334–334.

[67] E. M. Airoldi, D. M. Blei, S. E. Fienberg, E. P. Xing, Mixed membership stochastic blockmodels, Journal of machine learning research 9 (2008) 1981–2014.

[68] P. Yali, Y. Jian, L. Shaopeng, L. Jing, A biterm-based dirichlet process topic model for short texts, in: 3rd International Conference on Computer Science and Service System, Atlantis Press, 2014.

[69] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information processing & management 24 (1988) 513–523.

[70] C. C. Aggarwal, K. Subbian, Event detection in social streams, in: Proceedings of the 2012 SIAM international conference on data mining, SIAM, 2012, pp. 624–635.

[71] L. Prokhorenkova, Using synthetic networks for parameter tuning in community detection, in: International Workshop on Algorithms and Models for the Web-Graph, Springer, 2019, pp. 1–15.

[72] T. Yoshida, Toward finding hidden communities based on user profile, Journal of Intelligent Information Systems 40 (2013) 189–209.

[73] J. Yang, J. Leskovec, Defining and evaluating network communities based on ground-truth, Knowledge and Information Systems 42 (2015) 181–213.

[74] I. Inuwa-Dutse, M. Liptrott, I. Korkontzelos, Detection of spam-posting accounts on twitter, Neurocomputing 315 (2018) 496–511.

[75] I. Inuwa-Dutse, I. Korkontzelos, A curated collection of covid-19 online datasets, arXiv preprint arXiv:2007.09703 (2020).

[76] W. W. Zachary, An information flow model for conflict and fission in small groups, Journal of anthropological research 33 (1977) 452–473.

[77] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, S. M. Dawson, The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations, Behavioral Ecology and Sociobiology 54 (2003) 396–405.

[78] L. A. Adamic, N. Glance, The political blogosphere and the 2004 us election: divided they blog, in: Proceedings of the 3rd international workshop on Link discovery, 2005, pp. 36–43.

[79] J. Leskovec, A. Krevl, SNAP Datasets: Stanford large network dataset collection, http://snap.stanford.edu/data, 2014.

[80] J. Leskovec, K. J. Lang, 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.

[81] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Defining and identifying communities in networks, Proceedings of the national academy of sciences 101 (2004) 2658–2663.

[82] L. Danon, A. Diaz-Guilera, J. Duch, A. Arenas, Comparing community structure identification, Journal of Statistical Mechanics: Theory and Experiment 2005 (2005) P09008.

[83] A. L. Fred, A. K. Jain, Data clustering using evidence accumulation, in: Object recognition supported by user interaction for service robots, volume 4, IEEE, 2002, pp. 276–280.

[84] M. Girvan, M. E. Newman, Community structure in social and biological networks, Proceedings of the national academy of sciences 99 (2002) 7821–7826.

[85] X. Zhu, Z. Ghahramani, Learning from labeled and unlabeled data with label propagation, Technical Report, Citeseer, 2002.

[86] A. Lancichinetti, S. Fortunato, F. Radicchi, Benchmark graphs for testing community detection algorithms, Physical review E 78 (2008) 046110.

[87] P. J. Bickel, A. Chen, A nonparametric view of network models and newman–girvan and other modularities, Proceedings of the National Academy of Sciences 106 (2009) 21068–21073.

[88] B. Karrer, M. E. Newman, Stochastic blockmodels and community structure in networks, Physical review E 83 (2011) 016107.

[89] L. C. Freeman, Some antecedents of social network analysis, Connections 19 (1996) 39–42.

[90] R. I. Dunbar, The social brain hypothesis, Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews 6 (1998) 178–190.