Neurocomputing 289 (2018) 195–219

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Multi-view community detection with heterogeneous information from social media data

Antonela Tommasel*, Daniela Godoy

ISISTAN, UNICEN-CONICET, Campus Universitario, Tandil B7001BBO, Argentina

ARTICLE INFO

Article history:
Received 6 September 2016
Revised 3 November 2017
Accepted 5 February 2018
Available online 9 February 2018

Communicated by Prof. Zidong Wang

Keywords:
Community detection
Social networks
Multi-view learning
Social graph
Community structure

ABSTRACT

Since their beginnings, social networks have affected the way people communicate and interact with each other. The continuous growing and pervasive use of social media offers interesting research opportunities for analysing the behaviour and interactions of users. Nowadays, interactions are not only limited to social relations, but also to reading and writing activities. Thus, multiple and complementary information sources are available for characterising users and their activities. One task that could benefit from the integration of those multiple sources is community detection. However, most techniques disregard the effect of information aggregation and continue to focus only on one aspect: the topological structure of networks. This paper focuses on how to integrate social and content-based information originated in social networks for improving the quality of the detected communities. A technique for integrating both the multiple information sources and the semantics conveyed by asymmetric relations is proposed and extensively evaluated on two real-world datasets. Experimental evaluation confirmed the differentiated impact that each information source has on the quality of the detected communities, and shed some light on how to improve such quality by combining both social and content-based information.

1. Introduction

Social networking sites such as *MySpace, Facebook*, or*Twitter* attract millions of users, who everyday publish an enormous amount of content in the form of pictures, tweets, comments and posts. Social networks can be defined as a set of socially-relevant nodes connected by one or more relations. Nodes in such networks are not limited to people, but also represent other entities such as Web pages, journal articles or geographical places, amongst other possibilities. Users of networking sites are required to create profiles where users can describe themselves by sharing their age, locations, interests and picture, amongst other things. Generally, social networks allow users to create and read content, and establish social connections with other users whose nature and semantics might differ from site to site. For example, followee relations in *Twitter*, or friendship relations in *Facebook*. Although the technological features of the different social networking sites are similar, the cultures that emerge around them are diverse [3]. Most sites encourage the maintenance of pre-existing social networks, whilst others help strangers to create new connections based on shared interests. In this context, understanding users' needs arises as a critical issue [9]. Users' needs could be regarded as users' desire to obtain information, which could be further specified as long-term (interests) or instant (intends) user needs. Nonetheless, needs are often latent, so inferring them from the observed data might be challenging.

Social networks affect the way people communicate and interact. The pervasive use of social media offers research opportunities for analysing the behaviour of users when interacting with their friends [32], and how such interactions evolve over time [43], in terms of patterns of appearing and disappearing relationships. Unlike social connections formed by people in the physical world, social media users have greater freedom to connect with a wider spectrum of people for distinct reasons. The low cost of link formation might lead to networks with relationships of heterogeneous nature, origin and strength. For example, in *Twitter*, a user might follow others because they publish interesting information, they have the same interests, they are celebrities or popular individuals in the micro-blogging community, or only because they share some common friends, amongst other possible explanations. As a result, topological relations could lead to the existence of casual links, which could hinder the utilisation of algorithms solely based on topology. Hence, the nature of structural information must be carefully analysed in conjunction with other sources of information or data views to effectively assess the significance and importance of relations. In addition to social information

* Corresponding author.
E-mail addresses: antonela.tommasel@isistan.unicen.edu.ar (A. Tommasel), daniela.godoy@isistan.unicen.edu.ar (D. Godoy).

https://doi.org/10.1016/j.neucom.2018.02.023
0925-2312/

indicating friendship or simpler user interaction, there are other information sources that might implicitly define connections between users in social media. For example, whether two users use the same terms, hashtags, or post on the same topics. It is worth noting that the content users consume or post might depend, for example, on their mood and environment [9]. In light of the fact that users' needs are implicit, comprehensive research is needed for discovering the mapping between the heterogeneous, and possible multimedia, information in social networks and users' needs, and how such mapping can be enriched with contextual information.

One fundamental problem in social networks is the identification of groups of users when group membership is not explicitly available. A group, or community, can be defined as a set of elements (users, posts or other elements) that interact more frequently or are more similar to other community members than to outsiders. Community detection has proven to be valuable in diverse domains such as biology, social sciences and bibliometrics. For example, community detection techniques can be used for identifying groups of users with similar purchase history enabling the creation of more efficient recommendation systems that could better guide customers and enhance business opportunities as in *Amazon* [16], for detecting topics in collaborative systems [25], for identifying real-world landmarks in *Flickr* by clustering photos [26], for detecting events on *Twitter* streams [1], for matching high-quality answers to questions in the context of a question answering system [11], or for solving the influence maximisation problem in *Foursquare* [19].

Several techniques for community detection can be found in the literature. However, most of them only focus on one data view, even though neither social relations nor content by themselves can accurately indicate community membership. For example, in *Twitter* social relations might be extremely sparse and two users might belong to the same community even if they are not explicitly socially related. Conversely, social media content might be topically diverse and noisy for extracting valuable topical-based relationships. Combining multiple data views as required by social media data poses new challenges. For instance, how to integrate the different views by adequately assessing their importance in the social network, or how to determine whether such integration could actually improve the quality of detected communities.

Considering the increasing amount of information available in social networks and the necessity of integrating heterogeneous data, this paper focuses on the needs and challenges of combining multiple information sources for performing community detection. This work studies how to integrate multiple social and content-based views or information sources aiming at improving the quality of the detected communities. The final goal of the paper is to provide some insights on how to select the relevant views to consider for the task to develop according to the characteristics of the network under analysis. It is worth noting that the selection of the views to integrate depends on the elements available on the social network under analysis, such as the characteristics and semantics of social relations, the semantics of the messages users' exchange, or the content of such messages, amongst others. Moreover, several alternatives are proposed for integrating the semantics conveyed by the edge directionality embedded on the selected views. Finally, an extensive experimental evaluation of the benefits of combining the different views on diverse social networking sites is performed.

The rest of this paper is organised as follows. Section 2 discusses related research. Section 3 defines the nature of the diverse views to consider in the analysis, and a technique for combining them, as well as exploiting the semantics of edge directionality. Section 4 describes the experimental evaluation performed over real-world datasets. Finally, Section 5 summarises the conclusions drawn from this study and presents future lines of work.

2. Related work

Generally, social networks are analysed by means of graphs, representing a group of nodes or vertices, which are connected by links or edges. Edges can be directed (as the Followee/Follower relation on *Twitter*) or undirected (as the friendship relation on *Facebook*). Communities refer to potentially overlapping groups of nodes that have dense connections within the community, but sparse connections with nodes of other communities. Communities can be defined globally or locally, depending on whether a reduced subset of nodes or the whole network is considered. According to graph theory [20], communities have also been defined as cliques (every node is adjacent to each other) or connected components (every pair of nodes is connected by at least a path). In this context, the goal of community detection techniques (also known as graph clustering techniques) is to divide the nodes into communities (or clusters), such that the nodes of a particular community are similar or connected in some pre-defined sense [30]. For example, in some cases it might be desirable to obtain communities of similar order and/or density. Interestingly, not all graphs present a structure with natural communities. In the case of a uniform graph structure in which the edges are evenly distributed over the set of vertices, the discovered communities will be rather arbitrary.

Community detection has proven to be valuable in a diverse set of domains. Thus, several techniques for community detection can be found in the literature. The effort has been recently concentrated on addressing the challenges posed by the heterogeneous nature of social media data by combining diverse social networks [7,24] or sources of information, such as social and content information [28,33,38,42], similarity and interaction patters [13,40], and social, content and user similarity [27,32]. The existing techniques do not only differ on the considered information sources, but also on how such sources are combined. Particularly, this Section reviews techniques based on conditional or probabilistic models [35,38,39,42], matrix factorisations [7,24,27,28,32,35], and matrix integration [33,40].

Many tasks, in addition to community detection, can benefit from the integration of multiple and heterogeneous sources. For example, Xu et al. [36] chose to combine topological information derived from users' interactions in a university through a virtual mobile network with content-based information extracted from user profiles. Call records were used to establish the topological relationships, which varied according to when the calls were made, how many calls were made and their durations. Additional information regarding the faculty to which the users belonged, the dormitory and roommates was also considered. On the other hand, [11,44] combined social and content-based information in the context of question answering systems. Zhao et al. [44] tackled the problem of expert finding. To that end, the authors combined both information sources by means of a graph regularised matrix completion method for estimating the missing values in rating matrices (based on content-based information) with the social relations amongst users.

Discriminative conditional models for combining social and content information were proposed in [38,42]. Yang et al. [38] applied a conditional model for social analysis including hidden variables to model the probability of a node to be linked with another, and a discriminative content model for diminishing the impact of irrelevant content features. Experimental evaluation was based on two citation networks, in which nodes corresponded to scientific articles, edges represented citations, and content was described by keywords. Similarly, Zhang et al. [42] proposed a probabilistic model combining node attributes and topological

information. Experimental evaluation was based on *Twitter* and *Facebook* datasets from SNAP.[1] Content features varied according to the analysed dataset. For the *Twitter* dataset, they were hashtags and mentions, whereas for the *Facebook* dataset, they were the information in users' profiles, such as home town, birthday and political associations. In both cases, optimisation was performed by means of Expectation Maximisation, resulting in models that outperformed state-of-the art techniques based on social links, content or combining both sources of information.

Similarly to the previously described works Wang et al. [35] and Yang et al. [39] proposed conditional models combining network topology and node semantic attributes for detecting overlapping communities. Both works did not only identify communities, but also semantically annotated them. Yang et al. [39] probabilistically modelled the interaction between network structure and node attributes, which allegedly helped to improve the robustness of the technique in the presence of noise in the network structure. The presented approach has a linear runtime regarding the size of the network. The network modelling aimed at capturing three intuitions. First, community affiliations influence the likelihood that nodes are connected. Second, the degree of such influence is different across communities. Third, each community independently influences the node connection probability. Such intuitions were regarded as a logistic model, deriving in a convex optimisation problem. Evaluation was based on five social networking sites (*Facebook, Google+, Twitter, Wikipedia* and *Flickr*). The defined node attributes depended on the network under evaluation. For example, for the *Wikipedia* network, attributes were defined in terms of the links to other articles, in *Flickr* they were defined based on the used photos' tags. In *Facebook* and *Google+*, attributes were defined in terms of users' gender, job titles and institutions, amongst others. Finally, in *Twitter*, hashtags were selected. Results showed that the approach outperformed topology-based, node attribute-based and hybrid methods in terms of accuracy, even in noisy networks. The highest performance differences were obtained for the *Wikipedia* and *Flickr* datasets. Finally, the semantic of communities was analysed for the *Facebook* and *Wikipedia* datasets. As regards *Facebook*, education-based attributes (such as "school name" or "major") were highly correlated with communities' semantics, whereas work-based attributes were not. On the *Wikipedia* network, the approach was able to detect thematically close communities.

Zhang et al. [41] proposed a unified framework combining user friendship network analysis with author-topic modelling. First, the analysis of the friendship networks generates a community distribution of users, which is then used as prior knowledge by the content analysis. In turn, this analysis produces a set of community topics and user authorities on those topics, by assuming that topics can be modelled as a multinomial distribution over words. Finally, the community and topic distributions are combined to compute the final community memberships of individual users. The combination was performed by a non-linear strategy in which the community membership of users is linearly proportional to the membership derived from their social network, and exponentially proportional to their topical interests. Experimental evaluation carried out on small-scale *Delicious* and *Twitter* datasets showed that the algorithm was able to discover meaningful communities and their topics in a unified way. Moreover, the discovered communities exhibited denser friendship connections and higher content similarity than communities obtained with state-of-the-art techniques.

On the other hand, Pei et al. [27], Qi et al. [28], Tang et al. [32] and Wang et al. [35] proposed combining multiple information sources based on optimisation functions to be solved by non-negative matrix factorisations. In this regard, Wang et al. [35] based their approach on defining the propensities of nodes to belong to communities. Evaluation was based on three real-world networks (*Citeseer, Coral* and *WebKB*) comprising scientific publications. In all cases, node attributes were defined as the terms included in each scientific publication or web page. The approach was compared to topology-based, node attribute-based and hybrid methods (including [39]). All baselines were outperformed by the approach, showing its adequacy for accurately identifying community structures. Nonetheless, the approach was not evaluated in the context of dynamic short-text social media data; hence, results might not be generalisable to such domain. The semantic analysis of the detected communities was performed based on a *Last.fm* dataset in which node attributes included the list of most listened music artists and tag assignments. According to the authors, when selecting the top 10 node attributes, communities were deemed as cohesive. However, the rationale for choosing only 10 terms was not clarified, and the analysis was manually performed, thus no semantic similarity metric was computed. Moreover, it was not explored the cohesiveness of communities when selecting more attributes.

Both Pei et al. [27] and Tang et al. [32] use on matrix factorisation to discover communities of users. Tang et al. [32] chose to concatenate all content-based information sources, and combine them with the social information. The joint optimisation problem requires computing several arithmetic operations between matrices, which could negatively affect the computational complexity and thus, its applicability on high-dimensional datasets and real-time applications. Experimental evaluation was based on both synthetic and social media datasets from *BlogCatalog* and *Flickr*, including tags and comments. Particularly, nodes represented users, connected by friendship links, whereas the content-based information comprised the tagging, commenting and reading activity. Results showed that the quality of the detected communities depended on the quality of the selected information sources, as integrating more data sources introduced noise and redundant information, reducing the quality of communities, while increasing the problem's dimensionality. The authors suggested to consider short texts as additional information, as proposed in this paper.

Pei et al. [27] combined not only topological and content-based information, but also message similarity and user interactions. Experimental evaluation was based on two small-scale *Twitter* datasets comprising politicians, and a dataset of scientific papers. In contrast to the previously presented works, results showed that techniques solely based on social information performed better than those based on content. The authors stated that social relations better captured user interests, whereas content information introduced noise. However, as the evaluation was based on datasets with strong social and politics relations, there is no guarantee that the assumptions would held on general-purpose datasets where social relations might respond to diverse reasons.

In contrast to the described works that exploited node content, Qi et al. [28] assessed edge content, which models specific information regarding the nature of relationships and interactions between users. The authors proposed an edge-induced matrix factorisation for embedding edges into a latent vector space based on social information. Experimental evaluation was based on the *Enron* e-mail dataset, and a dataset collected from *Flickr*. In both cases, nodes corresponded to users. In the former case, edges corresponded to the e-mails both users had exchanged including their content, and in the latter case, edges were created if both users had marked the same picture as favourite, including the tags of all images marked as favourite by both users. Results showed that content-based algorithms outperformed social-based algorithms, implying that content provides useful information. The algorithms combining social information and edge content performed better

[1] http://snap.stanford.edu/data/.

than those considering node content. However, combining edge content to represent the node content did not always improve results as it mixed the content information from diverse edges.

The described approaches refer to heterogeneous information extracted from a unique social networking site. Nonetheless, users can participate in multiple networks simultaneously, then, each social networking site could provide additional information to help unveiling information about the users, existing in the other networks. Thus, community detection techniques could leverage not only on heterogeneous information belonging to a single network, but also on information belonging to multiple networks. In this context, Nguyen et al. [24] and Comar et al. [7] leveraged on the fact that users have profiles and connections in different social networking sites for detecting communities. Nguyen et al. [24] collapsed the information of multiple social networks into a unique representation, proposing two alternatives to join the information belonging to multiple instances of the same node. The first alternative collapses multiple instances of a node into a unique one, whereas the second one connects matching pairs of instances by edges adopting different coupling schemas (diagonal, categorical, star and full). Both alternatives were based on non-negative matrix factorisation algorithms. Although the representation techniques allowed to improve the results of baseline algorithms, building the graph representations (even when considering small datasets) incurred in a much higher computational complexity than the baselines, which might hinder their application on real social networking data.

On the other hand, Comar et al. [7] proposed to compute the adjacency matrix of each involved social network, in combination with a matrix linking them. The analysis could also include prior information regarding the potential relationships between the communities in the different networks. Then, communities are found by minimising the distance between the linking matrix and the product of latent factors of the adjacency matrices of each network. Experimental evaluation was based on *Wikipedia* and *Digg* users, and showed that *Wikipedia* was potentially useful as an information source for improving the quality of detected communities in social networking sites. The authors highlighted the fact that their technique could be applied in networks generated from multiple social networking sites as well as networks derived from heterogeneous nodes of the same networking site, as long as links between nodes in the different networks can be established, and acknowledged the scalability issues that might hinder the applicability of the technique on networks with millions of nodes.

Finally, related to this work are the studies carried out by Tang et al. [33], Zalmout and Ghanem [40]. Tang et al. [33] defined a processing pipeline involving four components and three intermediate steps. First, given a network, a utility matrix is built. Then, the utility matrix is processed to obtain a set of structural features by selecting the top eigenvectors. Such eigenvectors are supposed to represent the interaction patterns that could indicate the community partitions. Finally, a clustering algorithm is applied to the selected structural features to finally detect communities. Four alternatives are analysed for building the utility matrices: latent space models, block model approximation, spectral clustering and modularity maximisation. Each of the steps could imply considering information belonging to a unique network dimension or information derived from the integration of diverse dimensions. Particularly, four integrations are analysed. First, network integration (the closest strategy to this work), i.e. treating all dimensions as one by computing the average interaction network. Second, averaging the utility matrices. Third, integrating the structural features by applying Principal Component Analysis (PCA) to the concatenated structural features. Fourth, combining the obtained community partitions by reapplying the clustering algorithm to the obtained individual partitions. Experimental evaluation was based on *YouTube* data, comprising five data dimensions: contact, co-contact, co-subscription, co-subscribed and favourite videos networks. Results showed that the best results were obtained when considering structural integration, followed by utility integration. Conversely, the worst results were obtained when considering network integration.

Similar to the network integration strategy proposed by Tang et al. [33], Zalmout and Ghanem [40] presented a generic methodology for aggregating multiple data dimensions to discover communities of users. The methodology combines similarity and interaction patterns between users, such as the usage of hashtags, mentions, URLs or conversation engagement. Relations are represented as individual similarity matrices that are normalised and added to build the final graph. Then, a traditional community detection algorithm is applied. Experimental evaluation was based on a *Twitter* political dataset. Results showed that hashtags and URLs performed better when aggregated, whereas conversation engagement resulted in poor community quality. Moreover, removing frequent hashtags or mentions improved community quality. Unlike [32], the authors stated that aggregating all relations performed better than considering them separately. Additionally, as in [27], the small-scale dataset only comprised specific-purpose content, thus hindering the generalisation of conclusions to general-purpose datasets. Finally, the authors did not provide any means to differentiate the importance of the different relations. Contrasting with our study, their approach involved manually choosing the number of communities.

Unlike the presented approaches, this paper focuses on social networks comprising social media posts and the users who have written them, i.e. the goal is to discover groups of related posts based on their content and the social relations between their authors. Interestingly, none of the presented approaches explicitly treated edge directionality, thus ignoring the semantics of such relations. This paper proposes to analyse the effectiveness of several strategies for conveying the semantics of directed social relations.

3. Community detection based on heterogeneous social information

The first step to apply a community detection algorithm is to define the information that is going to be available to the algorithm, i.e. the information on which the underlying graph structure will be built upon. When analysing social media, multiple and diverse graphs can be defined. Nodes can represent not only real people, but also diverse entities such as Web pages, journal articles, countries, neighbourhoods, or positions, amongst others [21]. For example, if the goal of the community detection process is to predict new social relations between users or the influence a user has on his/her neighbourhood, nodes in the graph would represent the network users [36]. On the other hand, if the task aims at discovering relations amongst tags in folksonomies, nodes would represent tags [25]. In addition, nodes could represent photos if the goal is to detect geographical landmarks [26]. As in [1,38], this work aims at detecting communities of related posts in social media, hence each node in the built graph represents a social post. The discovered communities can be sub-sequentially integrated in diverse learning tasks such as clustering, topic detection, classification, or even in a feature selection technique.

Fig. 1 presents the overview of the process for detecting communities by combining heterogeneous information, starting from the original data feed extracted from a social networking site, up to the community discovery. Social media networks allow users to create content and establish social relations with others. As a result, social media data can be defined as a heterogeneous network that comprises not only information in the form of social or friendship relations, but also other sources of information

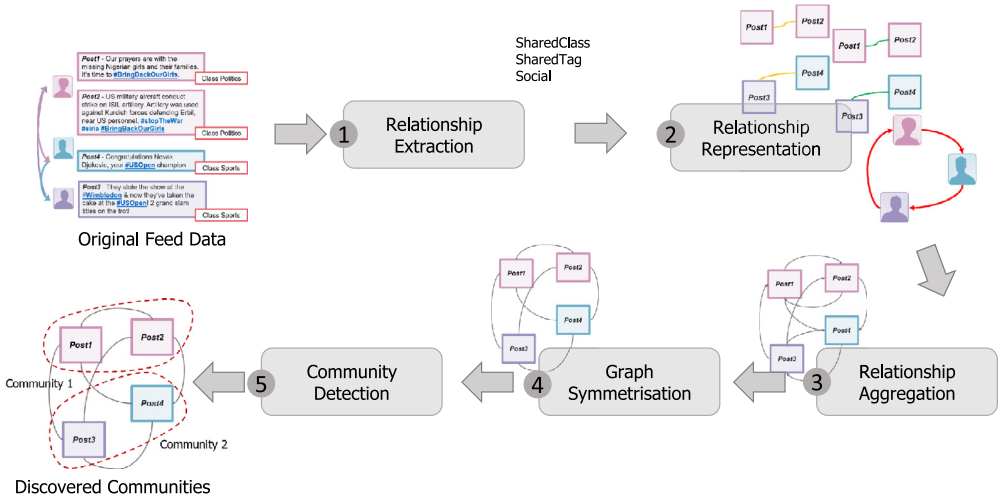


Fig. 1. Overview of the community detection process for heterogeneous information.

representing indirect connections between users or posts. For example, interactions between users or posts can originate in activities such as the interest of a user for a post expressed through the bookmarking of such post, or the frequency of comment and tagging actions, amongst other activities. These information sources provide different points of view of the same network, thus they can be useful for finding community structures. For example, in Steps 1 and 2 in the figure, three relationships are chosen to be analysed (namely, *SharedClass, SharedTag* and *Social*). The different types of relations need to be adequately leveraged when creating the graph representation of the network under analysis, as the result of Step 3 in the figure shows. In this regard, Section 3.1 presents different relations between nodes to be considered when creating the graph representation of the network.

The diverse relations established between nodes might embed also directionality information. For instance, when considering the Followee/Follower relationship in *Twitter*, the fact that user *A* follows user *B* does not imply the reciprocate (for example, in the resulting graph from Step 3, the relationship between *Post1* and *Post3* is not reciprocal). Even though relations might not be symmetric, most community detection techniques are based on the analysis of undirected graphs. In this context, Section 3.2 discusses several alternatives for convening directionality information in an undirected graph. Finally, once the heterogeneous relationships are analysed, the graph is created and symmetrised, and communities can be discovered (Step 5 in the figure).

3.1. Graph extraction

Most community detection techniques are purely based on the topology of the underlying social media network. However, in many applications, additional information that could help improve the quality of communities of social posts is either available or can be inferred. A distinct feature of social media posts is that they are potentially networked through user connections. For example, by considering the follower/followee or friendship links (i.e. social relations), several relations can be derived amongst the posts a pair of users have written [31]:

- Posts written by the same user are assumed to be related since they are more likely to belong to similar topics than randomly selected posts.
- If two users follow or are followed by a third user, their posts are more likely to have related topics than randomly selected posts.
- Posts are linked considering the friendship relations between their authors, i.e. a relation between two posts exists if the authors of such posts are connected in the social network. If there is a social link between users, they are likely to share interests, and thus, their posts are likely to be topically related.

In the context of social media data, both the graph topological structure (i.e. social relations between users) and node properties (i.e. posts characteristics) are important for improving the quality of the discovered communities. As a result, besides the social relations amongst posts derived from the actual social relations between their authors (i.e. post P_i is socially related to post P_j if its author is socially connected to the author of P_j), content-based relations could be defined amongst posts. The content resemblance or post categories (in case they are available) could also help to establish relations amongst them. Moreover, each microblogging site has specific characteristics and metadata that could be exploited for discovering meaningful relations between posts. For example, *Twitter, Instagram* and *Facebook* promote the usage of hashtags, which represent a type of label or metadata that aids in the search of messages of a specific theme or content. Additionally, *Facebook* allows searching for posts sharing specific activities, for example “listening Aerosmith” or “reading Oscar Wilde”. Posts containing the same hashtag or associated to the same activity can be assumed to be topically related. Fig. 2 exemplifies different types of complementary relations that could be observed between two nodes in a graph of posts.

For the purpose of this work, besides the traditional topological relation in which a link between two nodes representing posts exists if there are social relationships between users that published them, several content-based relationships between nodes were defined. Particularly, node content information is transferred to edges to characterise the specific relation between the linked nodes. By definition, all content-based relations are symmetric, i.e. they do not have directionality. Moreover, each relation could be assigned an individual scale-factor representing the importance of such relation in the final graph. Considering social networking sites that allow users to post content and tag it, relevant relations can be defined as follows:

- *Shared Tags.* An edge between two nodes exists if they share any tag (or hashtag). The weight of the edge is measured as the

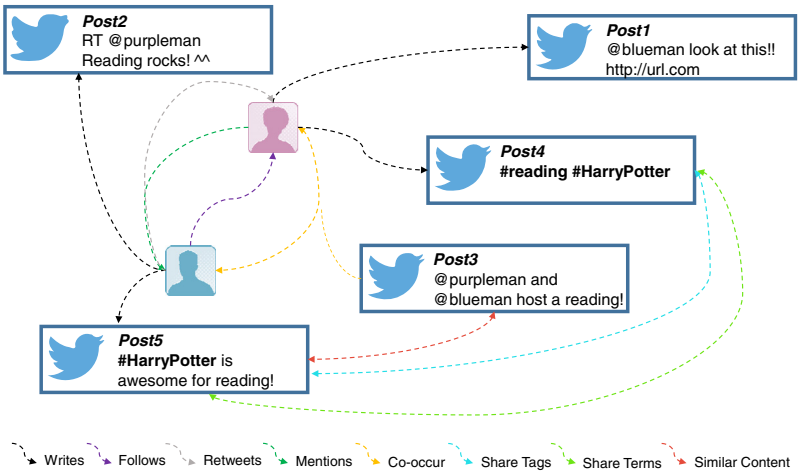


Fig. 2. Examples of possible links between posts.

percentage of shared tags amongst the total number of different tags comprised by the two posts.

- *Shared Class.* An edge between two nodes exists if they belong to the same class. All edges have a weight of 1. In those cases in which categories are organised in hierarchies or taxonomies (as in the *Open Directory Project*[2]), the edge weight could be computed as the distance between both categories.
- *Similar Content.* Measures the content resemblance of two nodes. A minimum similarity threshold could be imposed to avoid creating a complete dense graph. Thus, only edges with similarity above a certain threshold would be added to the graph. Diverse text similarity metrics could be adopted to define the nature and strength of similarities. For example, similarity could be expressed by simply computing the percentage of shared terms amongst the two nodes or by computing their Cosine Similarity.
- *Similar Comments.* As *Similar Content*, it measures the content resemblance of two nodes according to the Cosine Similarity between the comments each post has received.

Additionally, when considering social networking sites that allow users to tag or comment other user's posts, additional social relations could be defined to consider such interactions or social actions:

- *Tagged By Same Users*. Users can show interest in posts by tagging them. Then, posts that are tagged by the same users can be assumed to be topically related and to share a stronger connection that those that are tagged by disjunct groups of users. As a result, the degree to which two posts are tagged by the same users could denote an important relationship between them. The extent to which two posts are tagged by the same set of users is computed as the Jaccard Index.
- *Commented By Same Users*. Similarly, the activity of commenting posts also allow users to show their interest. Hence, posts that have comments written by the same users can be assumed to share a stronger connection than those with no commenters in common. Consequently, the extent to which two posts are commented by the same set of users could be used as a source of a new relationship between such posts, which can be measured by the Jaccard Index.

It is worth noting that social information and content-based relations offer complementary views of data, in this case, posts. Thus, no individual relation alone might be sufficient for accurately determining community memberships [32]. For example, social information might be sparse and noisy, while content-based information could be irrelevant or redundant, hindering the community detection process. Hence, it is important to combine the different types of relations for performing community detection in social networks.

Content-based relations could be used either to establish new relations between posts that are not socially related (named *Independent* graph derivation) or to reinforce the social relations already found amongst posts (named *Weighted* graph derivation). In the former case, social and content relations are assumed to be independent from each other, i.e. edges in the graph represent not only social links but also separated content ones. Hence, when considering both types of relations independently, two nodes might be connected even when there is no explicit social connection between them. In this graph derivation the different relationships are integrated by adding their corresponding matrices, as Eq. (1) shows, where A_{Rels} represents the aggregated adjacency matrix, *Rels* is the set of selected relationships and A_i are the adjacency matrices. Note that no differentiation is made between the social and content-based relationships.

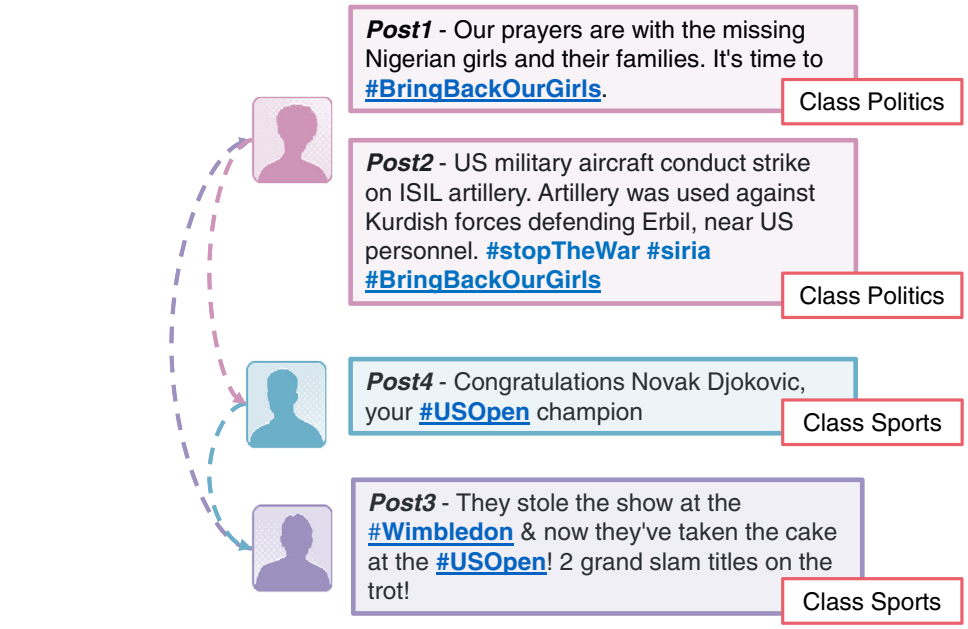
$$A_{Rels} = \sum_{i \in Rels} A_i \tag{1}$$

On the *Weighted* derivation, the graph only includes edges representing the social relation between nodes, whose strength or relevance is given by the content features. Thus, in this case, the quality of the social ties between nodes depends on an adequate definition of the content-based features, which should allow to fully exploit the social media data information. Eq. (2) shows how to compute the final adjacency matrix for this derivation, where A_{Social} represents the adjacency matrix for the *Social* relation and $Rels_W$ the set of relationships chosen for weighting the *Social* relationship. Note that this graph derivation also allows the integration of independent relationships, as showed by the second term in the equation. As it can be inferred from the equations, the computational complexity of the technique is of the order of $\Theta(n^{2*}v)$, where n represents the number of nodes in the graph (i.e. the number of posts) and v.

$$A_{Rels} = A_{Social} \circ \sum_{i \in Rels_W} A_i + \sum_{i \in \{Rels - Social - Rels_W\}} A_i \tag{2}$$

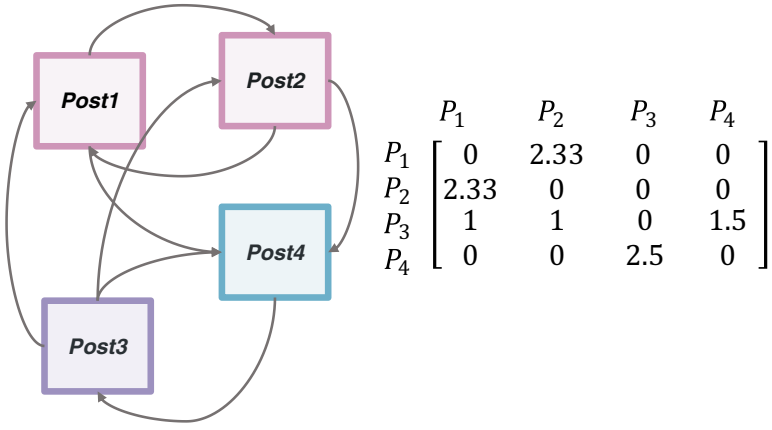
Fig. 3 presents an example of posts, the relations that could be established amongst them (*Social, SharedTag* and *SharedClass*), and how the final representation of the graph is derived from the inte-

[2] http://www.dmoz.org/.



(a) Original Data

(b) Detected Relations



$$\begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \end{array} \begin{array}{c} \begin{array}{cccc} P_1 & P_2 & P_3 & P_4 \end{array} \\ \begin{bmatrix} 0 & 2.33 & 0 & 0 \\ 2.33 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1.5 \\ 0 & 0 & 2.5 & 0 \end{bmatrix} \end{array}$$

(c) Final Graph Representation

Fig. 3. Multiple relations graph representation.

gration of such relationships (Fig. 3c). As it can be observed, Fig. 3b shows three graphs (which are accompanied by their matrix representation), each corresponding to one of the chosen relationships to analyse. The weight of the shown relations was defined as follows. In the case of the *Social* relations, a weight of 1 was assigned to the edge between two nodes (for example between *Post1* and *Post4*), when the author of a post followed the author of the other post (in this case, the author of *Post1* follows the author of *Post4*, as shown in Fig. 3a). In case the authors of two posts were not related (as the author of *Post3* with the other authors), the edge had a weight a 0, hence it was disregarded. For the *SharedClass* relation, a 1 was assigned to the edge if the two posts shared the same class (for example between *Post1* and *Post2* that, according to Fig. 3a belonged to class "Politics"), otherwise the weight was 0 (as in the case of *Post1* and *Post4*, which belonged to the "Politics" and "Sports" classes respectively). Finally, the score of the *SharedTag* corresponded to the percentage of shared tags between two posts. Considering the *SharedTag* relation between *Post1* and *Post2*, note that the two posts have three tags ("#BringBackOurGirls", "#stopTheWar", "#siria") out of which only one is shared by

the two posts ("#BringBackOurGirls"). Hence, the weight is computed as $1/3 = 0.33$. Similarly, *Post*4 and *Post*3 comprise two tags ("#USOpen", "#Wimbledon"), out of which only one is shared by the two posts ("#USOpen"), leading to the weight of 0.5. Then, Fig. 3c shows the final graph representation once all relations were aggregated into a unique graph by considering the *Independent* graph derivation. Note that the resulting graph is not symmetric, as for example, the relation between *Post1* and *Post4* is not reciprocal.

As it can be observed, the graph collapses multiple (and possibly heterogeneous) relations between two nodes into a unique edge, i.e. if multiple relations exist between two nodes, such relations are collapsed into a single edge. The weight of such edge would be equal to the sum of the weights of all the edges between nodes.

3.2. Graph symmetrisation

Once the relations between nodes are found, and the graph is built, the symmetric nature of relationships can be analysed. Generally, social relationships in social media data, as well as in other domains, are not symmetric, i.e. the fact that a user follows other user, does not imply that the second user reciprocates the relation. For example, the Followee/Follower relationships on *Twitter* or *Instagram* are not reciprocal.

While social networks exhibit diverse levels of reciprocity, most community detection techniques are based on the analysis of undirected (and perhaps weighted) graphs. Such techniques disregard the directionality of links, causing the loss of directionality information, thereby failing to accurately capture the semantics of the asymmetric relationships conveyed by the edges of a directed network [20]. Hence, the semantics captured by the relationships of undirected approaches substantially differs from the semantics of the directed relationships. Several works [12,29] have shown that the quality of the found communities could be improved by effectively taking into account edge directionality.

Two approaches can be applied for effectively including edge directionality in the community detection process. First, redefining the methods used for detecting the communities or assessing the quality of the detected communities. Second, applying transformations to directed graphs in order to attempt to retain the original graph semantics in an undirected graph. However, developing community detection techniques for directed graphs might be a difficult task [12]. For instance, a directed graph is characterised by asymmetrical matrices, so spectral analysis would be more complex. Moreover, whilst several graph concepts are theoretically well defined for undirected graphs (for example, density), they have not been extended to directed graphs [20]. Hence, only a few techniques can be easily extended from considering undirected graphs to consider directed ones. On the other hand, transforming the directed graph into an undirected one, i.e. symmetrising the directed graph, allows employing any of the algorithms or methods already defined for undirected graphs. This work explores several of the most common symmetrisation strategies available in the literature, which are described as follows.

3.2.1. Naïve graph transformation

This transformation ignores edge directionality and treats graphs as undirected ones. Although this is a common approach for handling directed graphs, it has several drawbacks that arise from the fact that the information represented by the directionality is ignored. First, the existence of data ambiguity. Naïve graph transformations introduce ambiguities and incorrect information in the graph, which do not represent the underlying semantic of the directed network. For example, assume that user A follows user B, but B does not reciprocate the relation. Using the naïve transformation each directed graph is replaced by an undirected one, thus a reciprocal relationship is introduced between users A and B, which adds an edge that did not exist on the original graph. Even when it can be argued that the new undirected edge could represent the similarity between users A and B, this does not always hold for both directions. For instance, user B could be a celebrity, whilst A could be just a devotee of B, thus mutual relationship and similarity might not actually exist. Second, deviations in the quality of the found communities. Even when the ambiguities could be ignore, they might still affect the final outcome of the community detection algorithm. In this case, communities that exist in the initial directed graph might not be identified in the transformed graph, leading to different results. This could be due to the fact that directed edges form interesting structural flow patterns and clusters.

3.2.2. Arithmetic-based transformation

In this case, the directed graph is transformed into an undirected one, whilst meaningfully capturing information and semantics about edge direction in the resulting graph. Then, community detection techniques designed for undirected graphs can be applied. Satuluri and Parthasarathy [29] analysed and proposed several techniques for transforming graphs based on arithmetic operations involving the adjacency matrix A of the graph. Particularly, two symmetrisation techniques are considered in this work. First, a simple symmetrisation in which the new adjacency matrix U, can be defined as $U = A + A^T$. This strategy is similar to ignoring edge directionality, except that in the case a pair of nodes is connected with edges in both directions, the weight of the edge in the symmetrised graph will correspond to the sum of the weight of the directed edges.

The symmetrised graph should be expected to include edges between nodes that share similar edges, but not including edges between nodes that do not share their connections. Although the simple symmetrisation is commonly used due to its simplicity, it might not be able to create edges between nodes that share connections but are not directly connected, as it only retains the same exact set of edges found in the original graph. In this regard, the second symmetrisation technique, the Bibliometric Symmetrisation, helps to cope with that situation. In this case, the new adjacency matrix is defined as $U = AA^T + A^TA$, where AA^T measures the number of common outgoing edges between each pair of nodes, and A^TA the number of incoming edges. The authors suggest to set $A = A + I$ before symmetrisating the graph to ensure that edges in the original graph are not removed.

Other more complex symmetrisation alternatives based on performing several arithmetic operations between matrices and diverse parameter tuning have been proposed by Satuluri and Parthasarathy [29]. As the complexity of arithmetic operations between matrices (particularly that of the matrix multiplication) is high, the techniques might not be useful in the context of high-dimensional social media data. In addition, parameters might be difficult to adequately tune in a high-dimensional and changing domain. Considering the high-dimensional domain in which the community detection technique will be implemented, this alternative was discarded for the purpose of this work.

3.2.3. Bipartite transformation

Directed graphs can be transformed into a bipartite undirected graph [15]. In the general case, nodes are placed on each partition according to whether they have outgoing or incoming edges. Particularly, the first partition of nodes contains every node that has outgoing edges, whereas the second partition contains every node that has incoming edges. According to graph theory [4], a natural correspondence exists between bipartite graphs and directed

Table 1
Twitter data collection main characteristics.

Number of instances	1036
Number of features	226,043
Number of classes	4
Number of following relations	251,522,840
Average number of followees	816
Average number of features per instance	1084
Average number of instances per class	259

Table 2
Flickr data collection main characteristics.

Number of instances	190,339
Number of textual features	947,829
Number of classes	81
Number of taggers	58,144
Number of commenters	569,765
Pairs of photos posted by the same user	77,909
Pairs of photos posted by users who are friends	8,825,738
Average number of features per instance	5
Average number of instances per class	1007

graphs, which can be easily modelled through the usage of the adjacency matrix. Let consider the adjacency matrix of the directed graph $A \in \mathbb{R}^{n \times n}$, where n represents the number of nodes in the graph. The adjacency matrix of the bipartite graph can be defined as: $B = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}$

As it can be inferred from the matrix definition, $B \in \mathbb{R}^{2n \times 2n}$, i.e. the nodes in the original directed graph are duplicated to avoid edges between nodes in the same partition. Note that in this case, the nodes in the different partitions represent the same type of elements.

4. Experimental evaluation

This section presents the experimental evaluation performed to assess the effectiveness of the proposed alternatives for leveraging on heterogeneous information provided by social media data, and is organised as follows. Section 4.1 presents the data collections used for evaluating the effectiveness of the presented technique. Section 4.2 presents implementation details and the metrics used for evaluating the different alternatives. Finally, Section 4.4 presents the results derived from the performed experimental evaluation.

4.1. Data collection

The performance of the technique was evaluated considering two real-world datasets. The first dataset was collected from *Twitter*[3] [45]. It included the content of more than 500,000 tweets belonging to 1036 trending topics, which were manually assigned to one of four categories: news, ongoing events, memes (trending topics that were triggered by viral ideas) and commemoratives (the commemoration of a certain person or event that is being remembered in a given day, for example birthdays or memorials). Table 1 summarises the main characteristics of the dataset. For the purpose of the experimental evaluation, each trending topic was considered as a node in the graph, i.e. each node grouped the tweet set associated to the corresponding trending topic.

The second dataset comprised data from the *Flickr* collection[4] as presented in [22], with the original images and metadata collected from the NUS-WIDE dataset[5] [5]. For each photo, the dataset included information regarding its owner, description, title, comments, tags, the groups in which the photo was posted and its manually annotated labels. Labels were considered as the category of photos, and hence the ground truth of the communities. In total, photos could be assigned 81 concepts. Concepts were extracted from frequently used tags in *Flickr*, representing either general concepts (e.g. “animal”) or specific concepts (e.g. “dog”), and they belonged to different general categories including scene, object, event, program, people and graphics. Only those photos containing at least one tag or description were kept. Additionally, the dataset provided information regarding edges between photos in *Flickr*, which allowed to infer the topological relations between the users and their photos. Such information included: the number of common tags, groups, and collections, an indicator for whether both photos were taken in the same location, an indicator for whether both photos were taken by the same user, and an indicator for whether the user that had taken the photo source of the edge was socially related to the user who had taken the other photo in the edge. The last two indicators were used to define the topological information of the network. For the purpose of the experimental evaluation, each photo was considered as a node in the graph. Table 2 summarises the main characteristics of the dataset.

4.2. Experimental settings

The Java programming language was chosen for implementing the technique. The graph implementation was based on that of the Gephi Toolkit[6]. The performance of all node relationships and symmetrisation strategies was evaluated considering the Gephi implementation of the Louvain algorithm [2]. Nonetheless, they could be used in combination with any other community detection algorithm or technique.

The quality of communities was evaluated by three types of scoring functions. First, functions that characterise the connectivity structure of a given community, built on the assumption that communities comprise sets of nodes with many inner connections and few outer connections. Considering the metrics presented in [18,37], a correlation analysis between the metrics’ results was performed according to the definitions and methods proposed in [8]. As data failed the normality tests, correlation was evaluated by the non-parametric Spearman Rank Order correlation. Results showed that metrics could be grouped in four groups, which were represented by *CutRatio, Density, FlakeODF* (Out Degree Fraction) and *Clustering Coefficient*. However, the results for three of the four groups did not showed significant differences amongst the different combinations of relations tested for the proposed datasets. Hence, only results of *FlakeODF* are reported. Second, a function characterising communities’ content cohesiveness: the average Cosine Similarity amongst all node pairs in the community (named*ContentCohesiveness*). Third, assuming the existence of class assignments in both datasets (the class of trending topics for the *Twitter* dataset, and the photo labels for the *Flickr* dataset), the entropy of the classes given the community assignments was also analysed.

To determine whether the graph size has an impact on the quality of the communities discovered by the proposed alternatives, different graphs sizes (ranging between 50 and 1000 posts) were considered in the experimental evaluation. For each graph size, five random partitions were generated. Then, for clarity of presentation, results across the different sizes were summarised by their mean value. For the *Twitter* dataset, the highest standard

[3] http://www.twitter.com/.
[4] http://snap.stanford.edu/data/web-flickr.html.
[5] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm.
[6] http://gephi.github.io/.

deviation on the scores was 0.03 for the *FlakeODF* metric. On the other hand, for the *Flickr* dataset, the highest standard deviation was 0.1 for the *Entropy* metric. In both cases, the deviation in the *ContentCohesivess* was similar (0.01 approximately).

Evaluation was performed considering both the social and content-based relations presented in Section 3. A social relation (named *Social*) between two nodes was established if the authors in a node followed authors of the other node. Each possible relation was evaluated individually and in combination with the others. Two variations of the *SimilarContent* relation were considered: a variation that created edges between every pair of nodes with a similarity greater than 0 (named *SimilarContent*), and one that imposed a minimum similarity of 0.6 for connecting two nodes (named *SimilarContent-0.6*). In the case of the *Flickr* dataset, an additional content-based relation and two social based relations were also considered: *SimilarComments, TaggedBySameUser* and *CommentedBySameUsers*.

Generally, the selection of a similarity threshold to deem two posts as similar depends on either specialists who fix a value, or trial/error processes, in which multiple values are tested until the result is satisfactory [10]. When thresholds are high, there is a risk of not finding interesting items, which in this particular case is represented by the significant content-based relations. On the contrary, low thresholds could find multiple irrelevant items. In this regard, threshold selection should be guided by the characteristics of the network under analysis, which would condition the distribution of posts' similarities, thus indicating the range over which posts similarities spanned. As a result, similarity thresholds could be defined based on the statistical distribution of similarities in the dataset. The selection of the statistical metric to guide the threshold definition is important, as it depends on the distribution type. Assuming the existence of outliers in the dataset, average measures of data cannot be used, as they do not give any indication of data dispersion. Instead, statistics that are not based on the supposition of a symmetric distribution of data, such as the interquartile range and outlier distribution, are needed. Outliers were detected using Tukey's method [34], setting $k = 1.5$ as suggested by the author. One of the advantages of the selected method is that it is applicable to both normal and skewed data since it does not make any distributional assumptions, and does not depend on the mean or standard deviation. Instead, it depends on the quartile definition.

When analysing the content similarity distributions for the different partitions of both datasets, it was found that most posts' similarities were concentrated on the lower scores, i.e. the similarity distribution was skew towards the left tail, indicating that most pairs of posts were not content related. Considering the skewed characteristics of these distributions, it could be assumed that as the values detected as outliers represent those values that are dissimilar to the majority of the values in the distribution, they would also represent the scores of those pairs of posts that could actually be deemed as similar. Hence, the similarity distribution was restricted to those scores that were marked as outliers. The restricted set of similarities was revealed to be more uniformly distributed than the original one. In this regard, the difference between the mean and the median scores was lower than the standard deviation, and no outliers were found for these distributions. Finally, the similarity threshold was defined as the average of the mean values found for each of the dataset partitions, i.e. 0.6. Interestingly, the same threshold was found for both datasets.

As exposed, the selection of the thresholds responded to the characteristics of the similarity distribution in the datasets, hence they cannot be directly generalised to different datasets. In case of analysing another dataset, the particular thresholds can be computed by the proposed methods. Note that the statistical properties of the defined threshold could be further explored aiming at optimising its selection.

As scores are computed for each individual community, they are averaged to obtain the score corresponding to a given community partition. Interestingly, several combinations of the defined relationships resulted either in only a single community containing all nodes in the graph, or in as many communities as nodes, i.e. each node had its own community. In this context, results are only reported for those alternatives finding a meaningful number of communities, i.e. a number between 1 and the number of nodes. Additionally, to ensure metrics' comparability, all results were normalised to the range [0; 1], and adjusted so that the highest scores represent the best ones.

4.3. Baselines for comparison

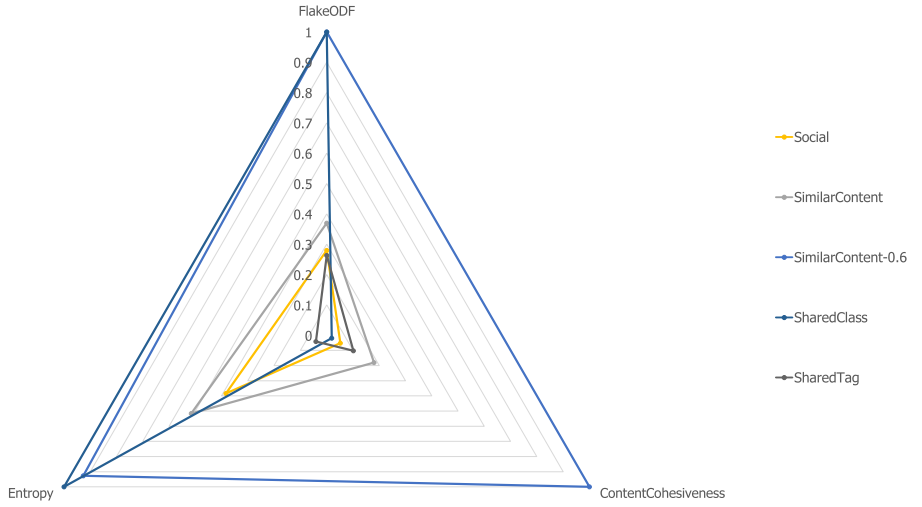
The presented approach was compared to several state-of-the-art techniques. Particularly, the experimental evaluation considered the alternatives in [40] (named *Zalmout and Ghanem*) and [33] (named *Tang* et al.). The same *Twitter* and *Flickr* datasets were used for this evaluation. A few considerations were made. Regarding *Zalmout and Ghanem*., the relationships to consider were selected according the two dimensions chosen by the authors (i.e. interaction and similarity dimensions). Nonetheless, considering that the approach cannot be directly mapped to the setting where our proposed technique is designed for (i.e. Zalmout and Ghanem [40] focused on networks of users, whilst this work focuses on networks of posts), and that the authors did not explicitly defined the considered data dimensions, the chosen set of relationships to analyse differs from the original paper. Second, the same implementation of the Fast Greedy [6] community detection algorithm was used, which was based on python's Igraph[7]. Third, as the selected algorithm generates a full dendrogram, the original approach required the definition of the number of communities to choose. For this particular evaluation, the number of communities was set to 6, 8 (the number of communities achieving the best results on the original paper) and the optimal community partition number based on optimising modularity.

As regards *Tang* et al., the performance of the four integration alternatives was compared to the presented approach. In the case of the structural feature integration, the smile[8] library was used for computing the Eigenvectors and PCA. Utility matrices were built considering the optimisation of modularity. Similarly to *Zalmout and Ghanem, Tang* et al. cannot be directly mapped to the detection of communities of posts as our technique proposes. Hence, the network dimensions used for analysing results correspond to sets of relationships similar to the ones originally used by the authors, and the independent combinations of relations obtaining the best results for our technique. In relation to the selection of structural features, as the authors only provided absolute numbers of selected features, for this evaluation, the number of selected features was set to the 10%, 50% and 100% of the total number of features.

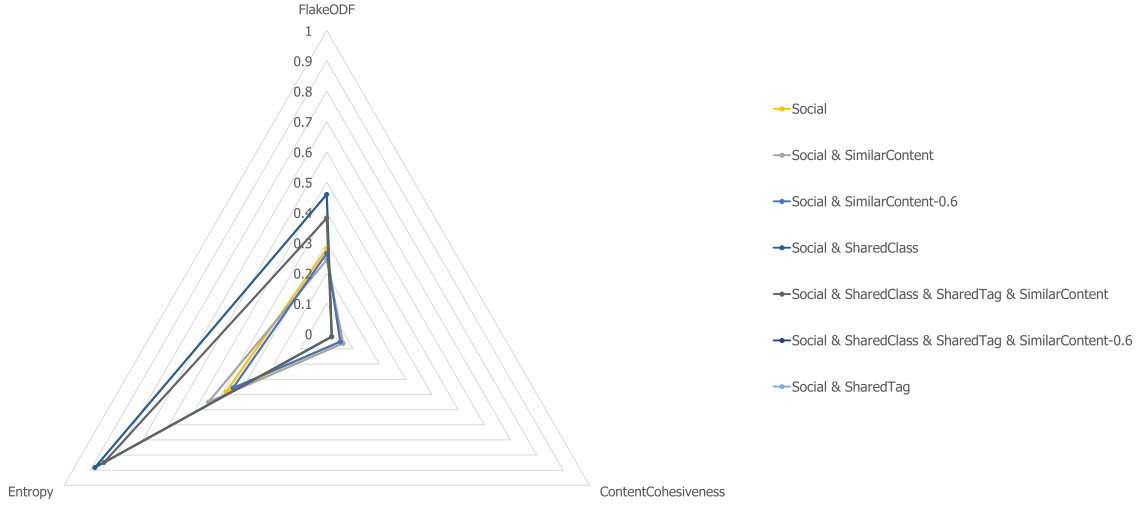
4.4. Experimental results

This section presents the results obtained for the evaluated datasets. For each dataset, three evaluations were performed. First, the importance of each independent node relationship was studied. Second, the effect of weighting the social view with the content-based relations was explored. Finally, the importance of the symmetrisation alternatives was analysed.

[7] http://igraph.org/python/.
[8] https://github.com/haifengl/smile.



(a) Individual Relationships



(b) Combination of Relationships

Fig. 4. *Twitter* dataset results – independent social and content views.

4.4.1. Results for the Twitter *dataset*

For this dataset, each node in the graph represented one of the manually classified trending topics. In this context, nodes could belong to one of the following categories: news, ongoing events, memes and commemoratives.

Independent social and content views

Fig. 4 shows the obtained results for the different combinations of node relationships using the Naïve symmetrisation. In general, the combination of relationships did not achieve neither high *FlakeODF* nor *Entropy* results. When individually assessing the defined relationships (Fig. 4a), the content-based views obtained communities of higher quality than the *Social* view. Particularly, all content-based relations allowed improving the *FlakeODF* results. As regards *Entropy*, the content-based relationships also obtained better results than *Social*, meaning that only considering the friendship relations between authors is not enough for identifying communities containing posts belonging to the same category. This could imply that the interests of users are not limited to only one category, and thus, they might publish posts belonging to diverse categories or connect with users posting on diverse categories.

As regards the *ContentCohesiveness* of communities, only *SimilarContent-0.6* found high quality communities, followed by *SimilarContent*, meaning that content-based relations could also introduce noise if not carefully analysed, and thus highlighting the importance of imposing a minimum threshold of similarity for regarding two nodes as content-related. The *Social* view allowed finding communities with a higher *ContentCohesiveness* than *SharedClass*, meaning that the category of posts is less representative of posts' content than users' friendship relations. Nonetheless, the *SimilarContent-0.6* and *SharedClass* views achieved similar *Entropy* results, implying that whilst the content of a post is related to its class, the class of a post is not sufficient to determine its content. Particularly, posts are divided into four categories (news, commemorative, memes and ongoing events) that do not represent actual posts' topics, i.e., two post could belong to the same category but contain unrelated content.

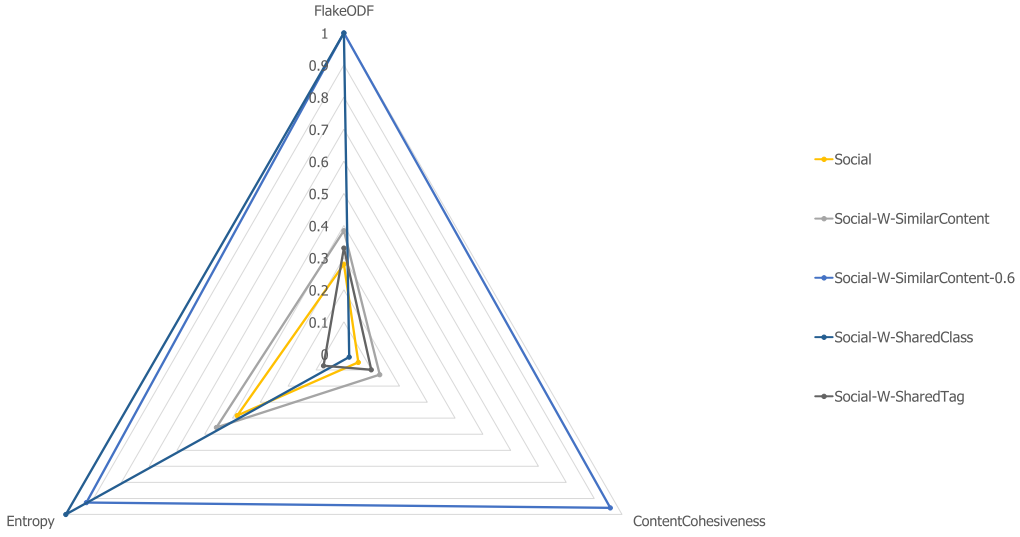
As it can be observed in Fig. 4b, the combination of *Social* and content-based relationships decreased, in most cases, the quality

Table 3
Twitter dataset results – ranking of best performing node relationships (independent social and content views).

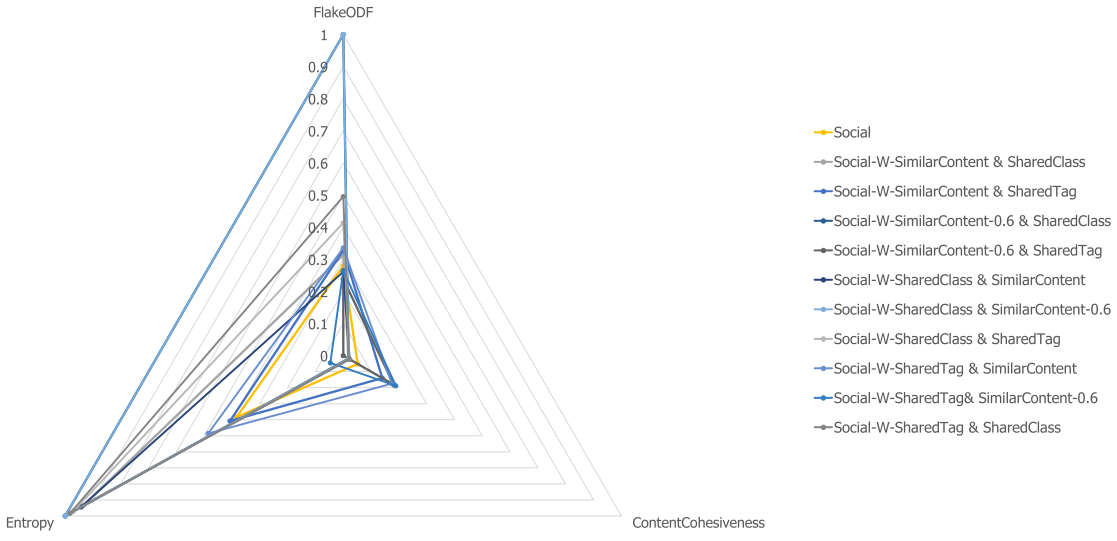
1. SimilarContent-0.6
2. SharedClass
3. Social & SharedClass
4. Social & SharedClass & SharedTag & SimilarContent-0.6
5. Social & SharedClass & SharedTag & SimilarContent

of communities obtained with respect to the content-based relationships alone. Although combining *Social* with *SharedClass* or any variations of *SimilarContent* improved the quality of communities regarding the simple *Social* view, the quality of communities was inferior to that of the individual content-based relations. Hence, it could be inferred that communities in *Twitter* might be guided for content-based relations, rather than for the social connections between users. As previously mentioned, the heterogeneous nature of social relations could introduce noise, hindering the identification of high quality communities. Interestingly, combining four of the defined relations (i.e. *Social, SharedClass, SharedTag* and any of the *SimilarContent* variations) obtained similar results to that of only combining the *Social* and *SharedClass* views. This might indicate that the information provided by *SimilarContent* is disregarded in presence of the *SharedClass* view. The difference between the *SimilarContent* alternatives remains noticeable across the *FlakeODF* and *Entropy* results.

The effect of the edge weighting is shown when comparing the *ContentCohesiveness* results of the *SimilarContent* view individually or in combination with the *Social* view. In the former case, such relation allowed to obtain content cohesive communities. In the latter case, however, the content cohesiveness of communities was diminished. This could be explained by analysing the absolute weight of edges. By definition, each relation weight is con-



(a) Individual Relationships



(b) Combination of Relationships

Fig. 5. *Twitter* dataset results – weighted social view.

Table 4
Twitter dataset results – ranking of best performing node relationships (weighted social view).

1. Social-W-SimilarContent-0.6
2. Social-W-SharedClass
3 .Social-W-SharedClass & SimilarContent-0.6
4. Social-W-SimilarContent-0.6 & SharedClass
5. Social-W-SharedTag & SharedClass

strained to the same range. However, the *Social* edges are prone to have higher weights than the content-based edges, thus being more important. For example, a *Social* relation between two posts would have a value of 1 in case the authors' posts are socially related, and 0 otherwise. In the case of the content-based relations, although their weights could be 1 (as when two posts belong to the same category), for the *SimilarContent* that would indicate that the posts have the exact same content, which is highly unlikely. As a result, the community detection algorithm is mostly guided by the *Social* view instead by the content-based ones. Hence, reinforcing the importance of adequately weighting the combined relations to optimise the quality of the discovered communities.

Table 3 ranks the node relationships that obtained the highest quality community partitions. The ranking was performed by averaging the results of all evaluation metrics for the Naïve symmetrisation strategy. All the ranked alternatives improved results of simply using the social relation. Excepting for the *SimilarContent-0.6* and *SharedClass* relations that were shown to diminish the quality of communities when combined with the *Social* view, the other content-based relations improved their results. For example, the *SharedTag* view obtained the best quality communities when combined with both *Social* and other content-based relations. Thus, the claim that information pertaining to a unique source offers a limited view of data is reinforced. These results allowed to conclude that introducing and combining content-based information is crucial for improving the quality of communities.

Weighted social view

Fig. 5 shows the results for the Naïve symmetrisation strategy and combinations of node relationships. For clarity reasons, "*Social-W-*" indicates that the social information was weighted with the content-based relation immediately named. As it can be observed, weighting the *Social* view with the content-based ones achieved similar results to the independent content-based views. For example, the results of *Social-W-SimilarContent-0.6* and *SimilarContent-0.6* are alike. However, weighting *Social* with *SimilarContent-0.6* or *SharedClass* improved the results of their independent combination. For instance, *Social-W-SimilarContent-0.6* improved the results of *Social* & *SimilarContent-0.6*. These results further emphasise the importance of adequately combining social information with other information sources to improve the quality of the detected communities. Moreover, using *SharedTag* for weighting the social information improved the results of considering it independently from the underlying social information. Additionally, an adequate weighting

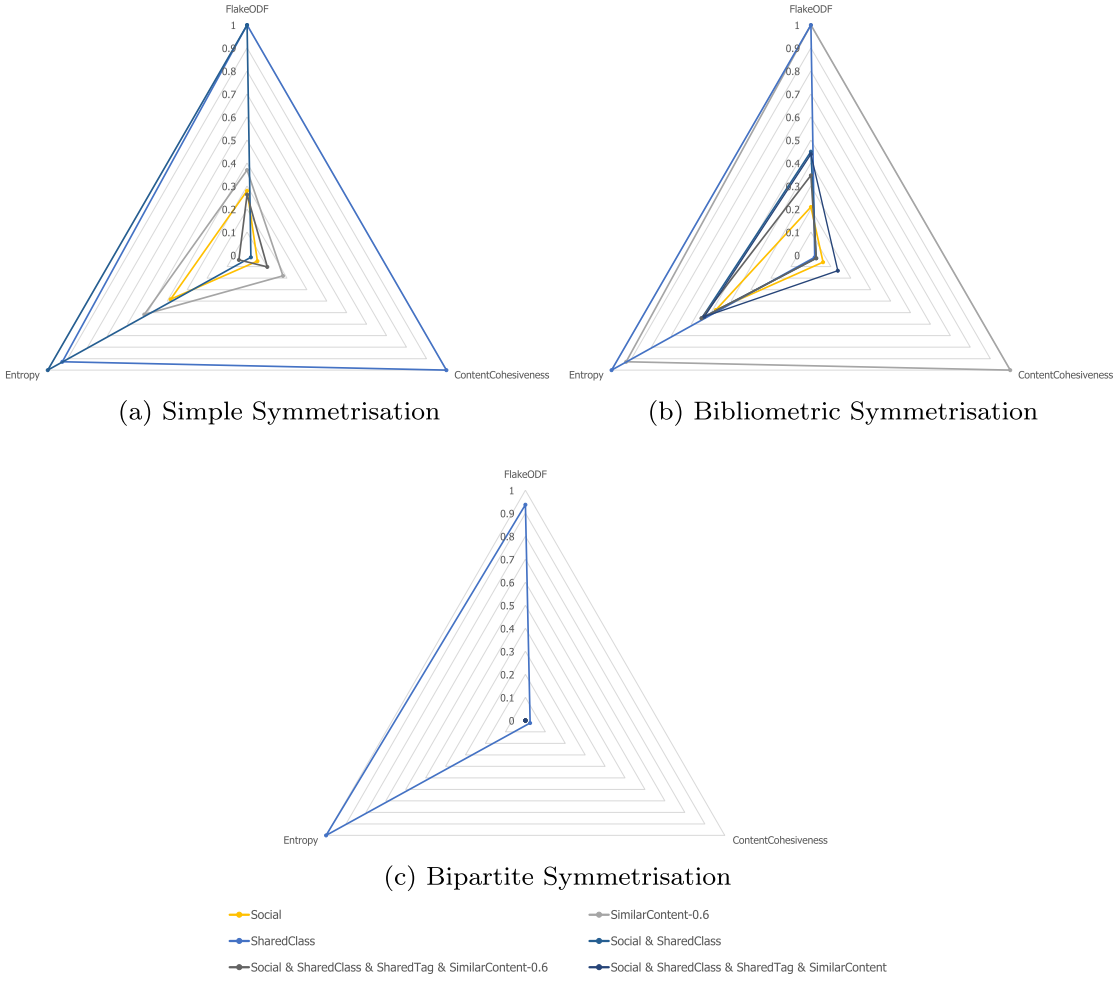


Fig. 6. *Twitter* dataset results – effect of the symmetrisation strategies on the independent relations.

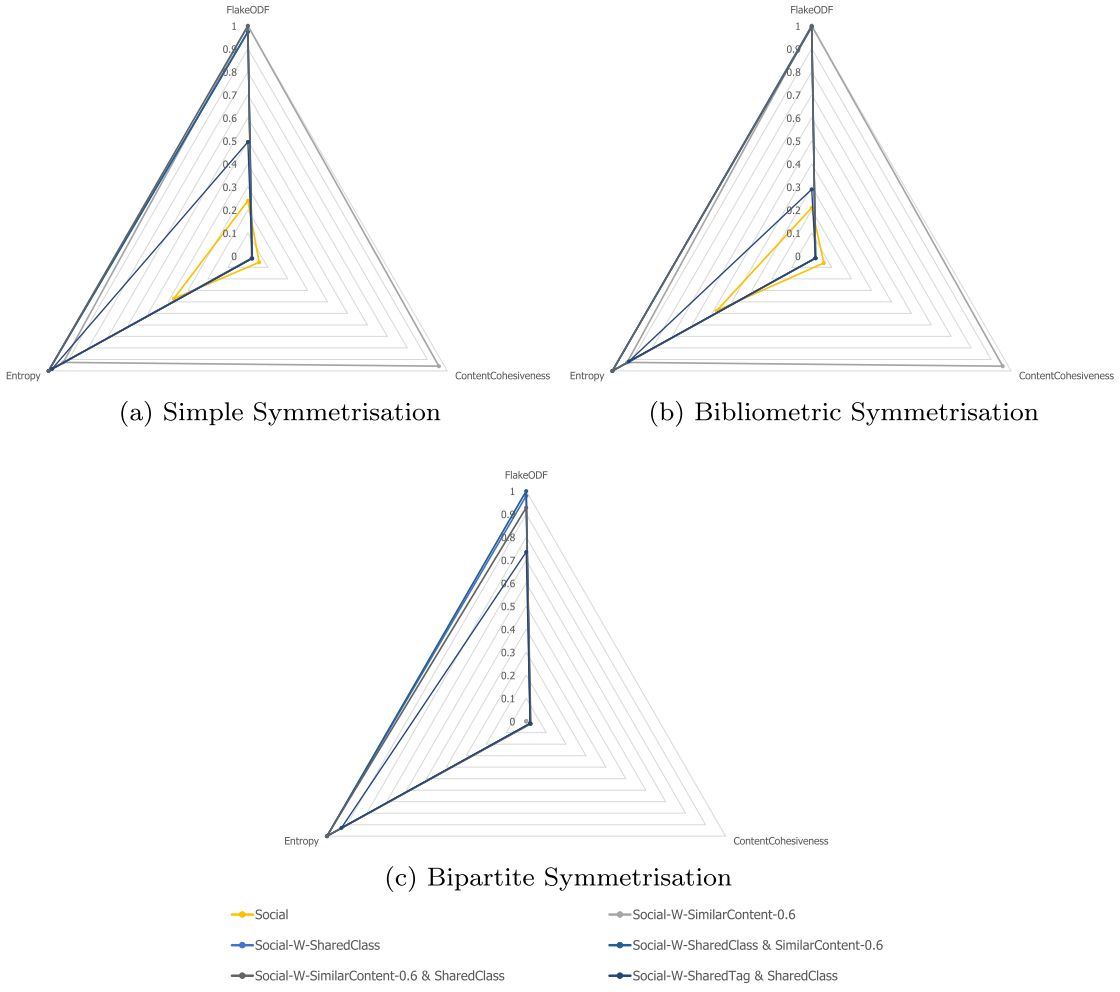


Fig. 7. *Twitter* dataset results – effect of the symmetrisation strategies on the weighted relations.

of the *Social* view with the*SimilarContent-0.*6 information allowed improving the *ContentCohesiveness* results of *Social*.

Regarding the combination of a weighted relation and other content-based views (for example *Social-W-SharedClass* &*SimilarContent-0.6* in Fig. 5b), results were better than when independently combining them (as shown in Fig. 4b). The *FlakeODF* and *Entropy* of communities was improved, i.e. communities were both more structurally and topically cohesive. Specifically, the *Entropy* of communities was high when including the *SharedClass* view. However, even though the *ContentCohesiveness* of communities was improved, results were still lower than when solely using the content-based views.

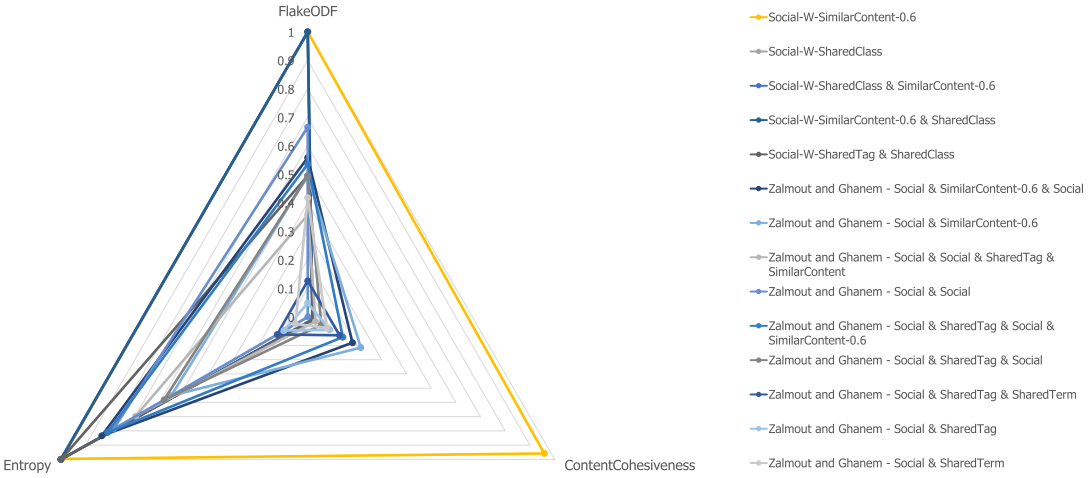
Table 4 ranks the node relationships that obtained the highest quality community partitions. In all cases, the weighted alternatives outperformed the results of only the *Social* view. Moreover, the best results were achieved when combining both the weighted social information with another content-based relation. Interestingly, all of the best ranked relations include either *SimilarContent-0.6* or *SharedClass*, further highlighting their relevance for finding high quality communities. These results reinforce the importance of content-based information for community detection in social networks. As when assessing the relations independently, the worst results were obtained when considering *SharedTag*.

Effect of the symmetrisation strategies

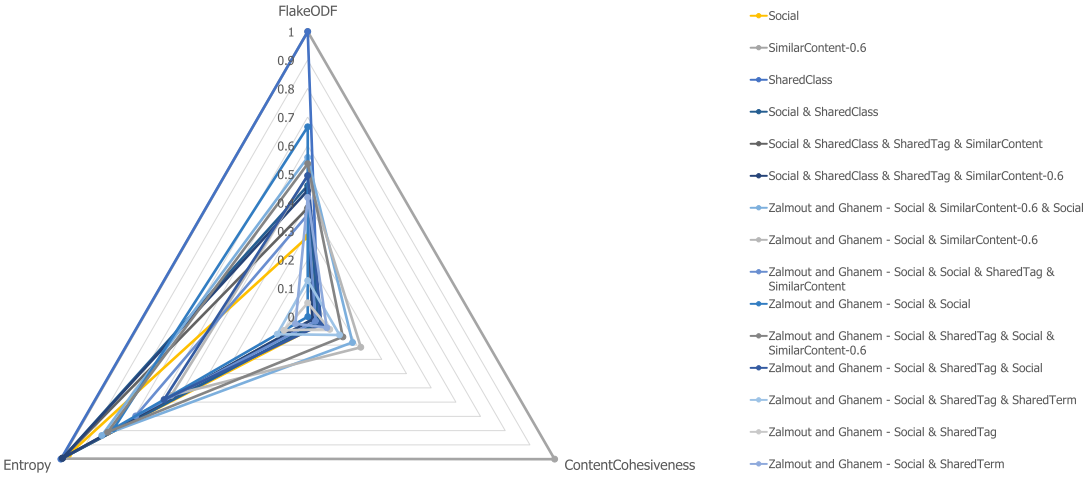
Figs. 6 and 7 analyse the effect of the presented symmetrisation alternatives on the best performing relationships listed in Tables 3 and 4, respectively. The most interesting results were those of the Bipartite symmetrisation, which resulted in the lowest number of reported node relationship combinations. As the Bipartite symmetrisation imposes a duplication of nodes, the structural composition of the graph changes, even when considering undirected relations such as the content-based ones. Consequently, the communities obtained based only on content relationships also changed. It is worth noting that, excepting when combining the *Social* and the *SimilarContent-0.6* relations (which was not one of the best performing strategies for the Naïve symmetrisation strategy), all other combinations were unable to find communities. Interestingly, the *SimilarContent-0.6* relation by itself did not found a representative number of communities. A similar effect is observed for the Bibliometric symmetrisation, which reduced the *FlakeODF* of communities, thus reducing the quality of the obtained communities in comparison with the other strategies.

The highest differences between the Simple and Naïve alternatives were found for *FlakeODF* in favour of the Simple symmetrisation. These results allowed inferring that the semantics conveyed by the directionality of social relations, when adequately assessed, can help to improve the quality of communities. Thus, it is not only important to select the node relationships to combine, but also the symmetrisation strategy to use.

As Fig. 6 shows, results for each symmetrisation strategy are similar to that of considering the independent combination of node relationships. As in the previous case, the Simple symmetrisation



(a) Independent Social and Content Views



(b) Weighted Social View

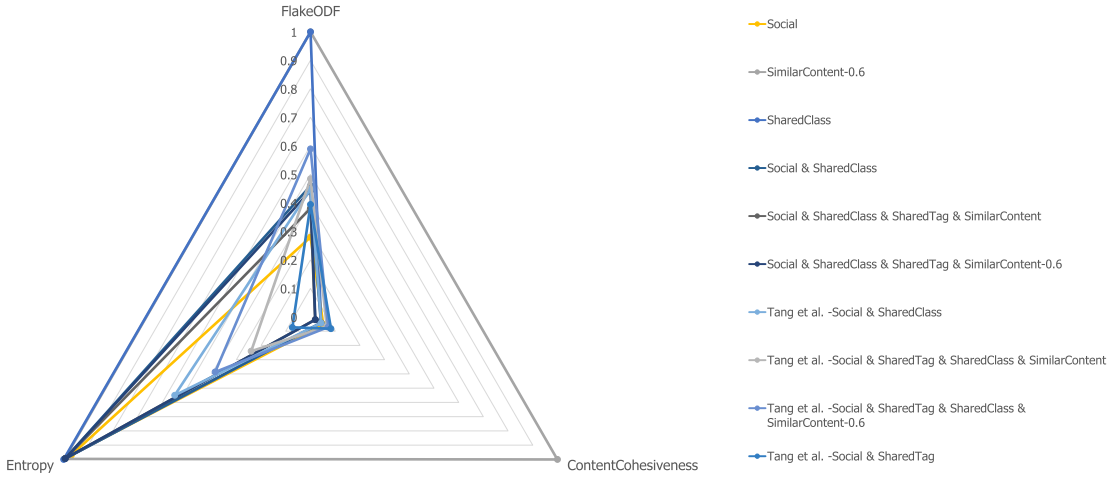
Fig. 8. *Twitter* dataset results – comparison to Zalmout and Ghanem [40]'s approach.

technique allowed finding the highest quality communities, improving the *FlakeODF* of the communities found by using the Bibliometric symmetrisation.

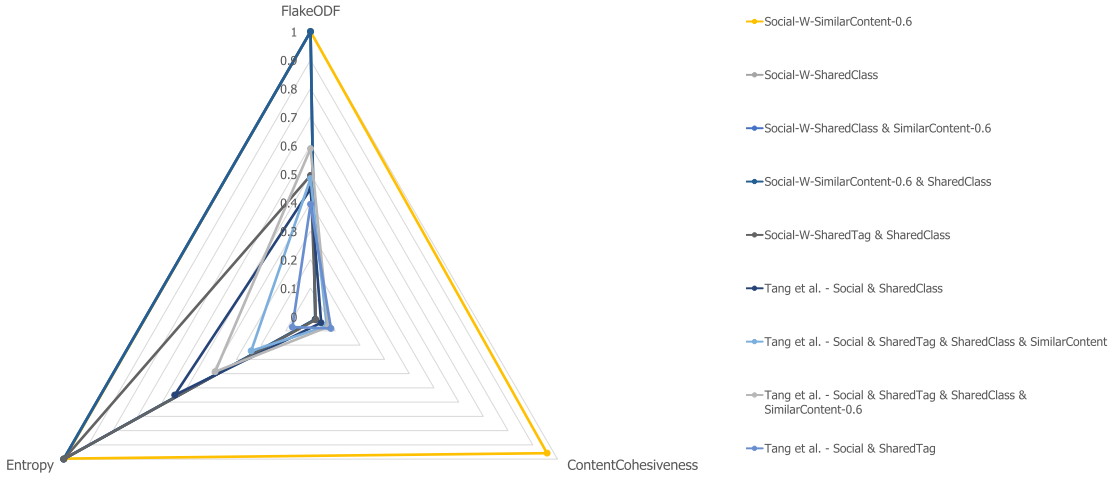
Comparison to state-of-the-art techniques

Fig. 8 compares the results obtained for the best combination of relationships for the *independent* and the *weighted* graph derivation with those of *Zalmout and Ghanem* for the *Twitter* dataset. As regards the number of communities, the best results were found when automatically selecting the number of communities that optimised modularity, instead of when fixing the number of communities as in [40]. In addition to those combinations of relationships only including a Jaccard Similarity assessment (e.g. *SharedTag* and *SharedTerm*) as defined in [40], the performance of *Zalmout and Ghanem* was also evaluated by the combinations of relations achieving the best results for our technique. Note that, for both graph derivations, our technique improves *Zalmout and Ghanem*'s results. As it can be observed, the communities found by *Zalmout and Ghanem* exhibited low *FlakeODF*, whilst achieving competitive *Entropy* (when analysing the *independent* graph derivation). Interestingly, only the best performing strategy of *Zalmout and Ghanem* (*Social* & *SimilarContent-0.6* & *SharedClass*) was able to outperform the *ContentCohesiveness* obtained by our technique. The worst results *Zalmout and Ghanem* results were obtained when combining *Social* & *SharedTag*, which were even worse than solely considering *Social*. The alternatives considering Jaccard Similarity (as in the original paper) performed worse than those considering Cosine Similarity, i.e. *SharedTag* and *SharedTerm* obtained communities of lower quality than *SimilarContent*. Regarding the *weighted* graph derivation, none of the *Zalmout and Ghanem* evaluated alternatives was able to outperform the results achieved with our technique. In average, the quality differences ranged between 90% and 1183% when considering the lowest and highest improvements for both graph derivations.

As regards *Tang* et al., the comparison of results is presented in Fig. 9. Considering the diverse thresholds, the best results were obtained when only selecting the 10% of the total number of structural features. Regarding the integration alternatives, similarly to the results in the original paper, the best results were obtained when integrating the structural features derived from each utility matrix. As it can be observed, *Tang et al.*'s results are similar to those of *Zalmout and Ghanem* in terms of *Entropy*. Both *FlakeODF* and *ContentCohesiveness* were lower than for *Zalmout and Ghanem*. The best results were obtained when combining *Social* & *SharedTag* & *SharedClass* & *SimilarContent-0.6*, closely followed by *Social* & *SharedClass*. Nonetheless, despite considering the same relations, our technique was capable of finding communities of higher quality. In average, the quality differences ranged between 186% and



(a) Independent Social and Content Views



(b) Weighted Social View

Fig. 9. *Twitter* dataset results – comparison to Tang et al. [33]'s approach.

445% when considering the lowest and highest improvements for both graph derivations.

4.4.2. Results for the Flickr *dataset*

For this dataset, each node in the graph represented one of the manually classified photos, comprising at least one tag or description. Each photo could belong to 81 different concepts, which represented elements visible in the photos. Note that photos could be assigned to more than one concept, which might not match the tag nor description created by the users.

Independent social and content views

Fig. 10 shows the results for the Naïve symmetrisation strategy. Unlike the results obtained for the *Twitter* dataset, most combinations of information sources achieved high *Entropy* and *FlakeODF*, implying that the communities found for this dataset are more strongly connected. Moreover, the quality of communities is higher than that achieved for the *Twitter* dataset.

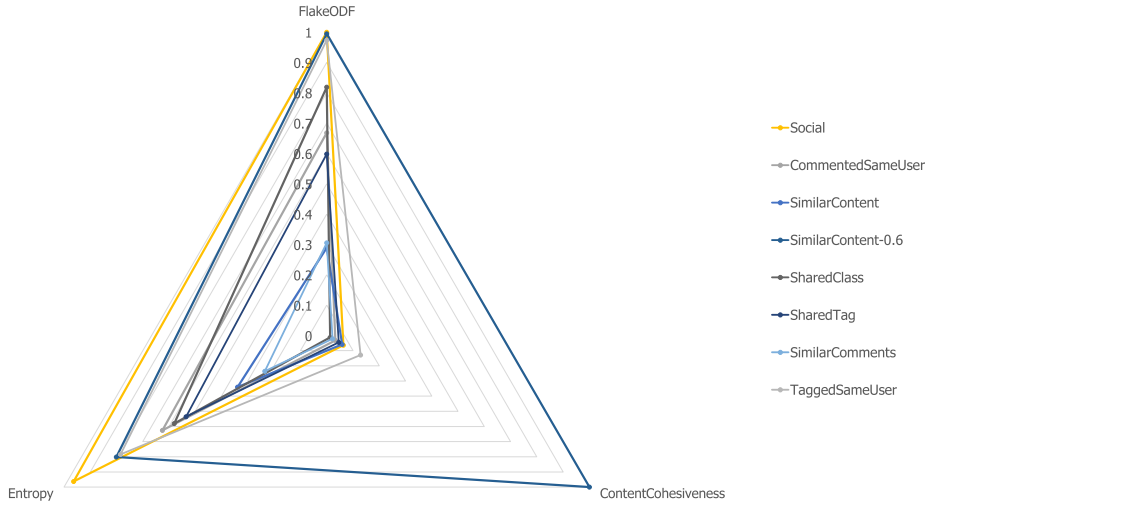
As regards the individual relationships (Fig. 10), only *SimilarContent-0.6* achieved relatively high *ContentCohesiveness*. In spite of creating a dense graph, the *SimilarContent* relation did not obtain neither high content nor class cohesiveness. Additionally, the *SimilarComments* or *SharedTag* views did not report neither highly content cohesive nor structurally connected communities.

Table 5
Flickr dataset results –ranking of best performing node relationships (independent social and content views).

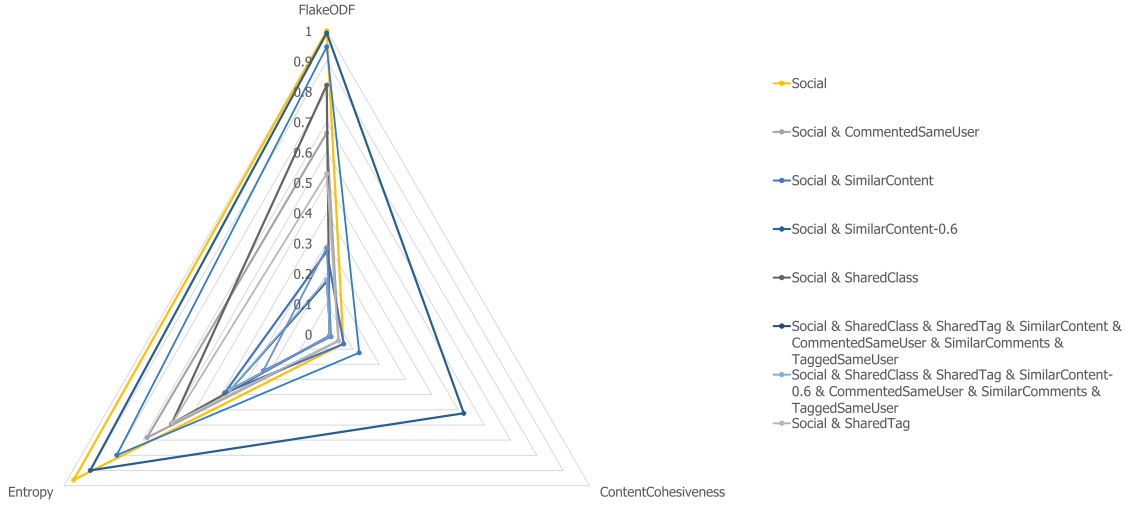
1. SimilarContent-0.6
2. Social & SimilarContent-0.6
3. Social
4. TaggedSameUser
5. Social & TaggedSameUser

As for the *Twitter* dataset, content and class cohesiveness results were not directly correlated, implying a dissociation between the content in the description and tags, and the label that was assigned to the photos. However, *SharedClass* did not achieve the best *Entropy* results, as it did for the *Twitter* dataset. These results imply that for this dataset, the labels assigned to photos are not sufficient for finding communities of photos belonging to the same category, thus continuing to expose the limitations of only using a single information source.

Individually considering the *Social* relationship achieved high *FlakeODF* and the best *Entropy* results. Thereby, it could be inferred that communities in *Flickr* might be guided by social connections between users' relations, rather than for content-based information. This could be related to the high degree of reciprocity of



(a) Individual Relationships



(b) Combination of Relationships

Fig. 10. *Flickr* dataset results – independent social and content views.

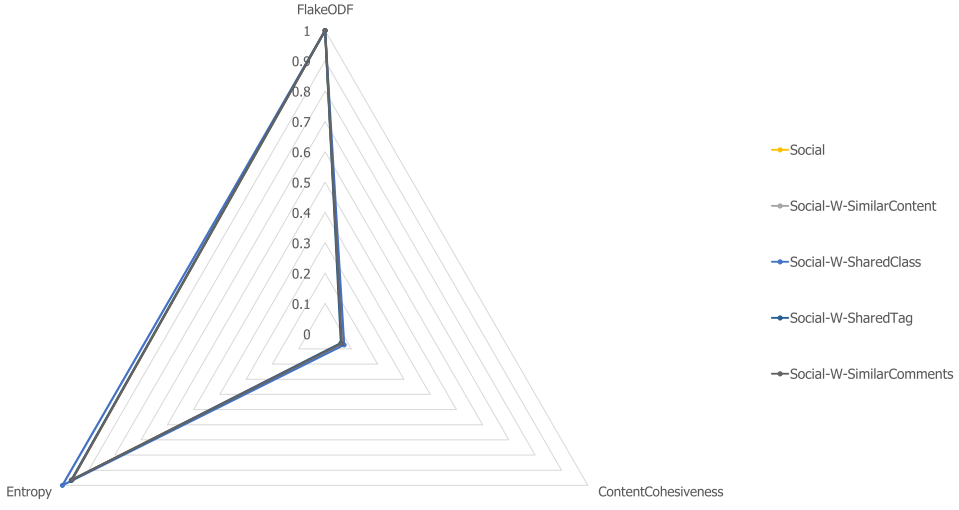
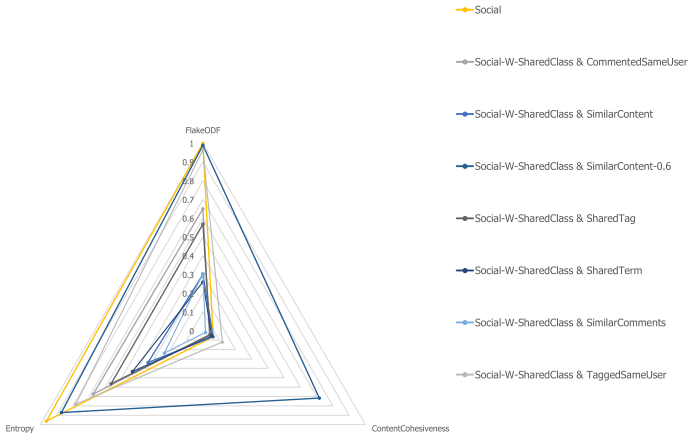
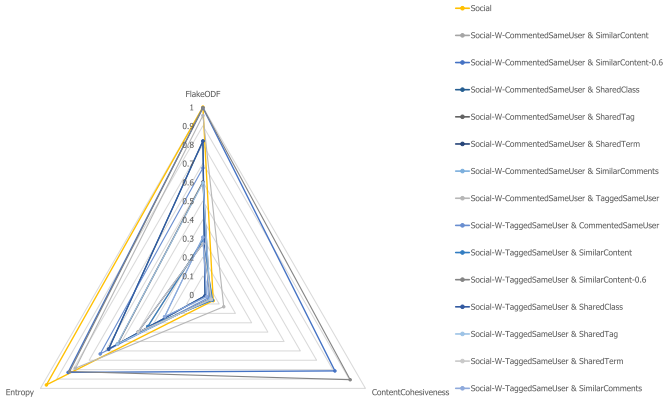


Fig. 11. *Flickr* dataset results – weighted social view – individual relationships.

(a) Combination of Relationships based on *SimilarContent*, *SimilarComments* and *SharedTag*



(b) Combination of Relationships based on *SharedClass*



(c) Combination of Relationships based on *CommentedSameUser* and *TaggedSameUser*

Fig. 12. *Flickr* dataset results – weighted social view.

Table 6
Flickr dataset results – ranking of best performing node relationships (weighted social view).

1. Social-W-TaggedSameUser & SimilarContent-0.6
2. Social-W-CommentedSameUser & SimilarContent-0.6
3. Social-W-SharedClass & SimilarContent-0.6
4. Social-W-SimilarContent & SimilarContent-0.6
5. Social-W-SharedTag & SimilarContent-0.6

social relations, which responds to the same characteristics of offline social relations [14], instead of showing content-based motivations. Finally, considering other topologically-based social relations (*CommentedSameUser* and *TaggedSameUser*) allowed to improve the content cohesiveness of communities whilst decreasing *Entropy* results. These results confirm the importance of the topological relations for this dataset.

When combining *Social* with the other content-based relations (Fig. 10b), as for the *Twitter* dataset, *SimilarContent-0.6* decreased its *ContentCohesiveness* results, whilst improving the *Entropy* of communities. Similarly, *CommentedSameUser* in combination with the *Social* view decreased both the *Entropy* and *ContentCohesiveness* of communities. Results showed that combining two topological relations was not as effective as individually considering them. As for the *Twitter* dataset, these results could be explained by the effect of edge weighting. Moreover, results support the claim that the diverse information sources might introduce noise, and thus, combining multiple relations might not always help to improve the quality of communities, as shown by the results obtained when mixing all information views, whose quality equalled to that of the worst performing individual content-based view.

Table 5 ranks the node relationships that found the highest quality community partitions. The ranking was performed by averaging the results of all evaluation metrics obtained for the Naïve symmetrisation strategy. As it can be observed, some of the best performing relations differ from those found for the *Twitter* dataset. Regarding the content-based relations, only *SimilarContent-0.6* appears amongst the best ranked strategies. The worst results were obtained by *SimilarComments* followed by the combination of all relationships. The results of *SimilarComments* could be explained by considering that, generally, comments in social media might be motivated by a desire of expressing opinions or sentiments, instead of describing the content they are commenting on. As a result, comments are not descriptive enough for the community detection task, as they would not help finding content nor class cohesive communities.

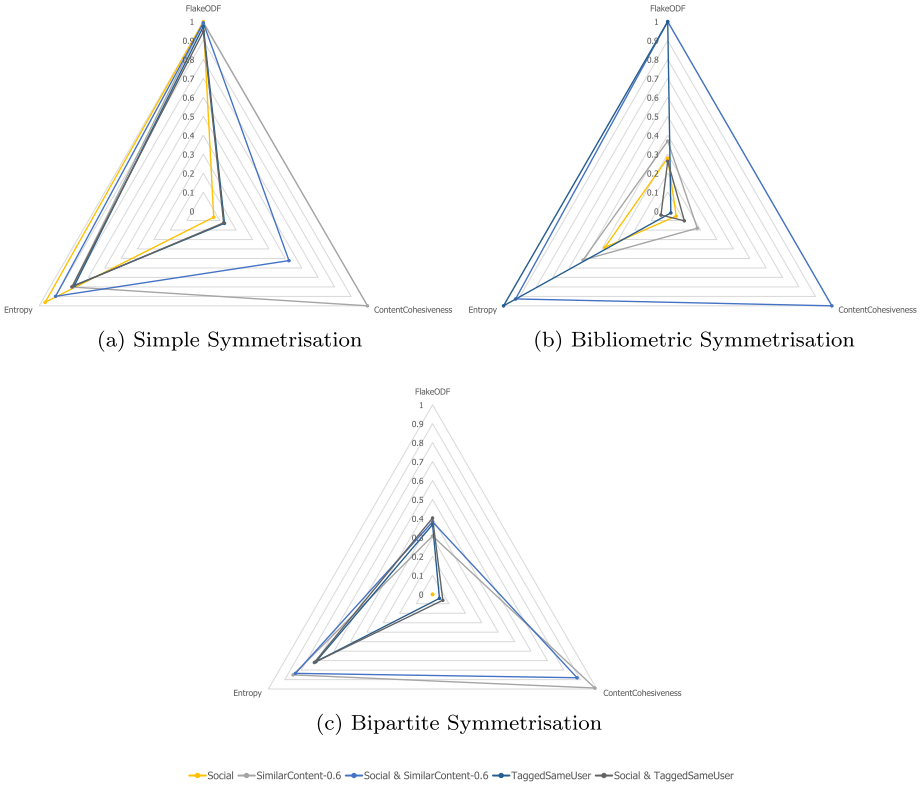
Note that the averaged results of *SimilarContent-0.6* are better than those of the *Social* view due to the improvements in content cohesiveness. Moreover, combining *Social* and *SimilarContent-0.6* improved the results of the *Social* view. As already exposed, *Social* appears as one of the best ranked relations. Also *TaggedSameUser* is one of the best performing relationships. These results further highlight the importance of the topological relations for this dataset.

Weighted social view

Figs. 11 and 12 show the obtained results for the Naïve symmetrisation strategy and the combinations of node relationships. As Fig. 11 depicts, weighting *Social* with the other defined relations caused all combinations to find community partitions of similar quality, which coincidently match the results of only using the *Social* view. The only exception was when using *SharedClass* as the weighting strategy, which improved the *Entropy* of communities. Interestingly, weighting *Social* with *SimilarContent-0.6* or *SharedTag* did not help to obtain a meaningful number of communities. As a result, none of the alternatives discovered communities with high *ContentCohesiveness*.

In the overall, weighting and combining the *Social* relation with the content-based ones (Figs. 12a–c) allowed improving the content cohesiveness of communities without reducing their *FlakeODF*. Moreover, *FlakeODF* results were also improved in comparison to the results of using the social and content-based views independently. As regards *ContentCohesiveness*, the best results were obtained when *Social* was weighted with the other topology-based relations (as shown in Fig. 12c) and combined with *SimilarContent-0.6*. Interestingly, weighting *Social* with *SharedClass* (Fig. 12b) did not achieved the best *Entropy* results. Instead, they were lower than when individually considering the *SharedClass* view.

Table 6 ranks the node relationships that found the highest quality community partitions. Unlike when assessing the diverse relations individually, the ranked alternatives outperformed the results of only considering the *Social* view. Moreover, the individual *Social* view obtained, in average, worse results than when weighting it with other relations. The two best performing combinations included weighting *Social* with the other topological-based rela-



(a) Simple Symmetrisation

(b) Bibliometric Symmetrisation

(c) Bipartite Symmetrisation

Fig. 13. *Flickr* dataset results – effect of the symmetrisation strategies on the independent relations.

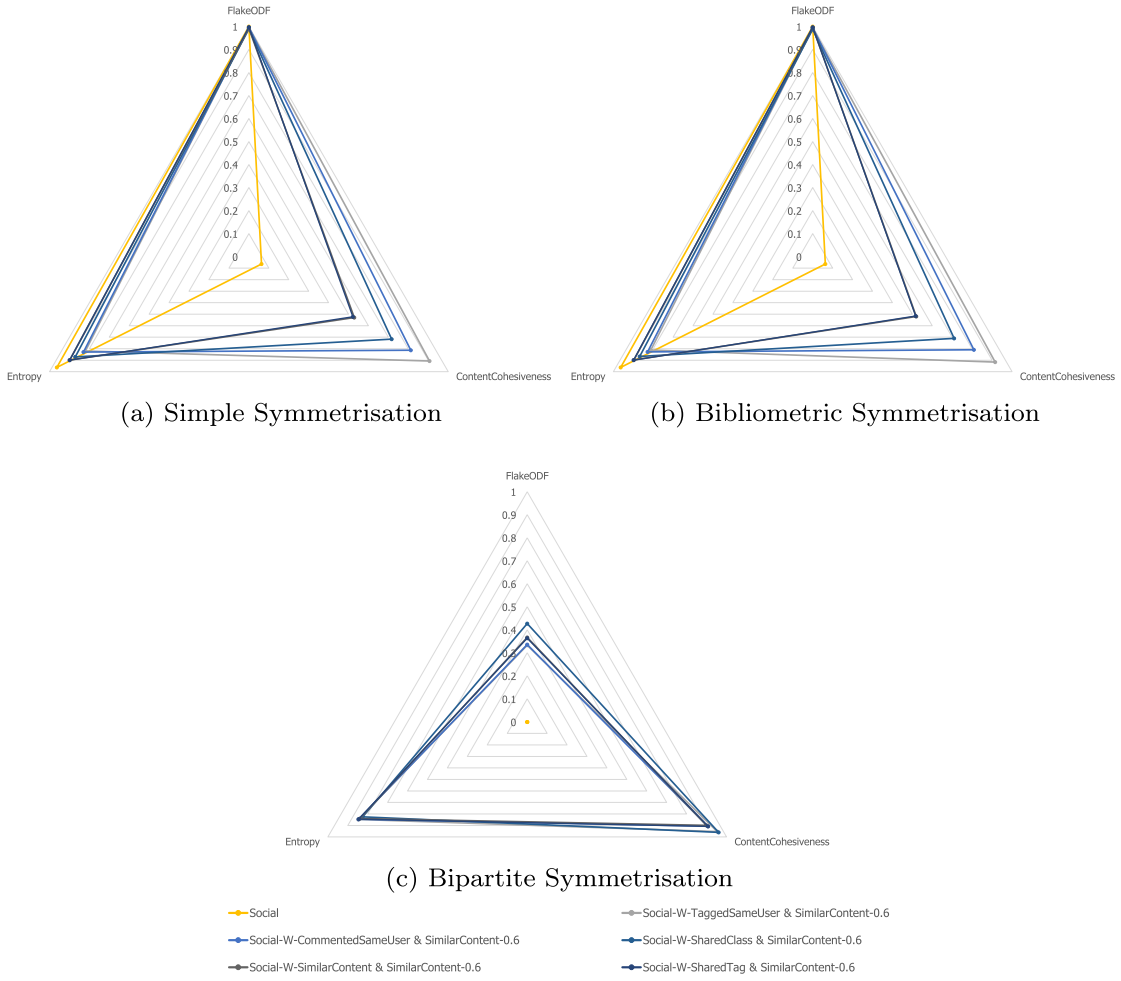


Fig. 14. *Flickr* dataset results – effect of the symmetrisation strategies on the weighted relations.

tionships, evidencing the importance of the topological relations for this dataset. Note that none of the individual weighted relations (shown in Fig. 11) ranked amongst the best performing ones. Instead, all the best ranked relationships include combinations with *SimilarContent-0.6*. Thus, it can be inferred that content-based information was also important for finding high-quality communities as a complement of other information sources. In addition, *TaggedSameUser, SimilarComments* and *SharedTag* were shown to achieve better results when used for weighting the *Social* view than when individually used. These results remark the positive effect of weighting the *Social* view on community quality, in contrast to individually considering the relations, which confirms the importance that the underlying social relations have on *Flickr*.

Effect of the symmetrisation strategies

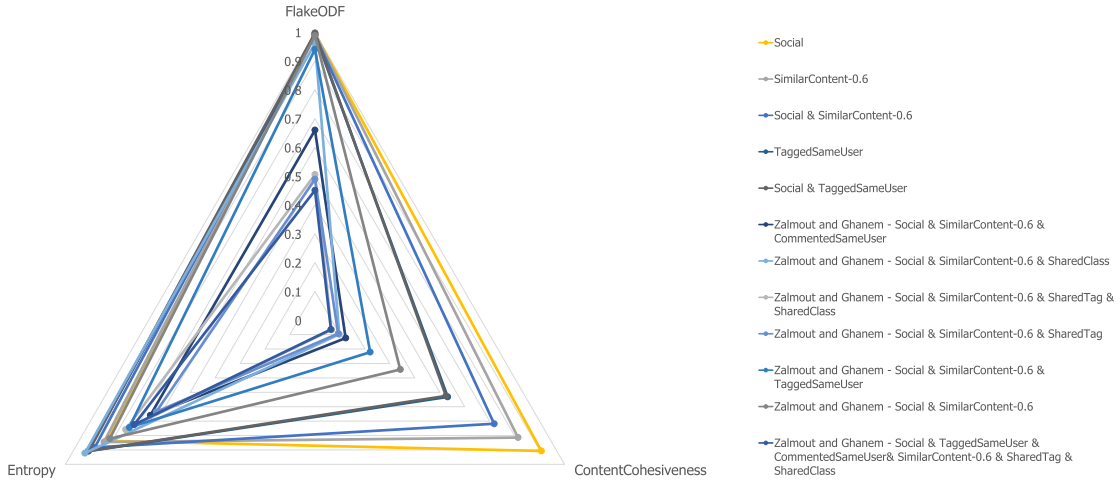
As for the *Twitter* dataset, Figs. 13 and 14 analyse the effect of the presented symmetrisation alternatives on the best performing relationships listed in Tables 4 and 6, respectively. The effect of the chosen symmetrisation alternative over the views' ability for detecting a representative number of communities was lower than in the *Twitter* dataset. In this case, only a few of the proposed combinations of relations did not found a meaningful number of communities for only one of the symmetrisation alternatives (Bipartite symmetrisation). Remarkably, one of such views is the *Social* one. These results confirm the differences between the diverse symmetrisation alternatives, and how they can affect the quality of the detected communities. Moreover, these results reinforce the importance of considering multiple information sources.

Similarly as when analysing the *Twitter* dataset, the Simple and Naïve symmetrisations obtained communities of similar connectivity. These results could be explained by considering the reciprocity degree of *Flickr*. Several studies [17,23] have shown that the reciprocity in *Flickr* is higher than the 70%. Consequently, the number of asymmetric relations is small when compared to the number of symmetric relations, which implied that the simple symmetrisation alternative did not contribute with new information, thus finding communities of similar quality. Finally, the *Bipartite* symmetrisation notably reduced the *FlakeODF* and *Entropy* of communities. As explained before, even though relations are symmetric the duplication of nodes imposed by the Bipartite symmetrisation, causes changes to the structural composition of the graph, which accounts for the differences in the quality metrics.

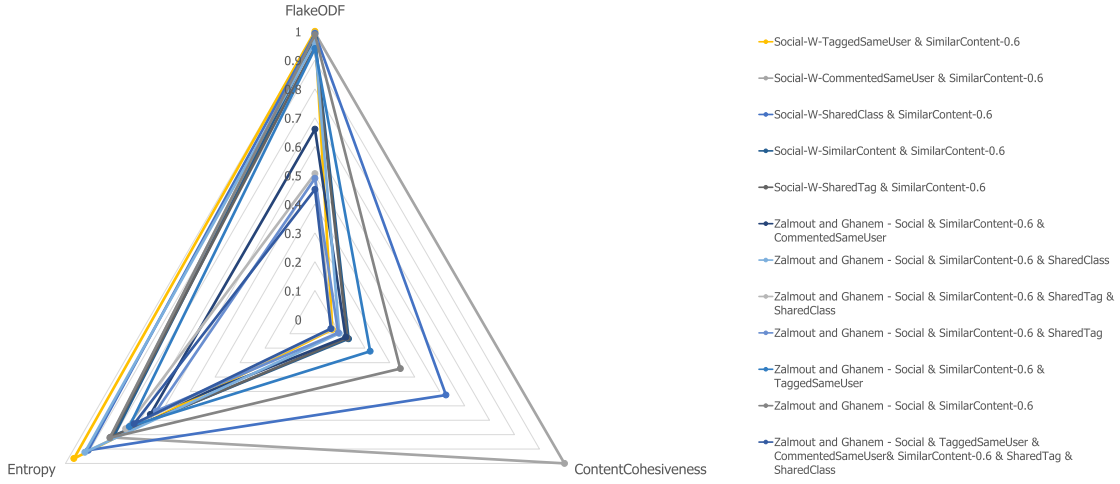
Despite lowering the structural cohesiveness of communities, the Bipartite symmetrisation in combination with the *Social* and *SimilarContent-0.6* views allowed to find the most content cohesive communities. These results highlight the importance of adequately assessing the content-based relations and their relevance for finding high-quality communities.

Comparison to state-of-the-art techniques

Fig. 15 compares the results obtained for the best combination of relationships for the independent and the weighted graph derivations with those of *Zalmout and Ghanem* for the *Flickr*



(a) Independent Social and Content Views



(b) Weighted Social View

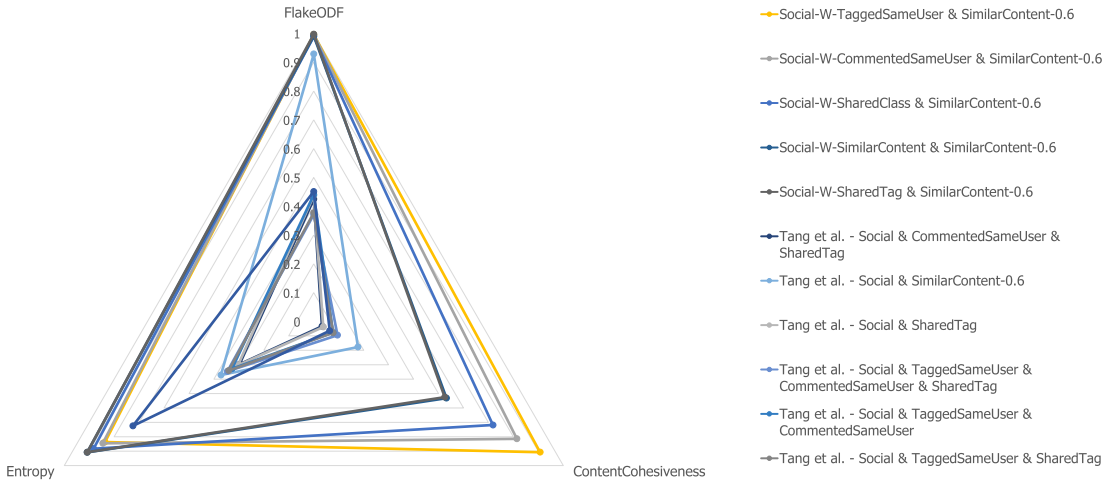
Fig. 15. *Flicker* dataset results – comparison to Zalmout and Ghanem [40]'s approach.

dataset. The best results were found when automatically selecting the number of communities that optimised modularity. Note that both graph derivations are capable of improving *Zalmout and Ghanem* results. As it can be observed, *Zalmout and Ghanem* achieved competitive results in terms of *FlakeODF* and *Entropy*. Small differences were found in favour of our technique for those two metrics. On the other hand, our technique obtained significantly better community partitions in terms of *ContentCohesiveness*, even though *Zalmout and Ghanem* was evaluated considering the *SimilarContent-0.6* relation. Nonetheless, in all cases, for all evaluation metrics, our technique outperformed every *Zalmout and Ghanem* results. Interestingly, the best *Zalmout and Ghanem* results were obtained when combining *Social* & *SimilarContent-0.6*, which is also the relationship combination that obtained the best quality partitions for our technique. In average, the quality differences ranged between 30% and 130% when considering the lowest and highest improvements for both graph derivations.

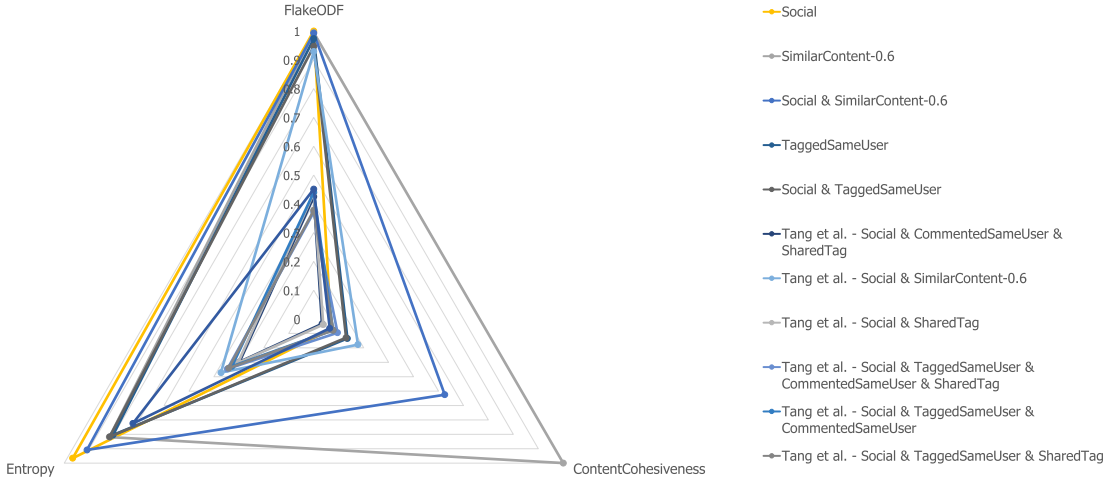
As regards *Tang* et al., the comparison of results is presented in Fig. 16. It is worth noting that not every combination of node relationships, integration strategies and threshold for selecting structural features could be evaluated due to the lack of convergence of the Eigenvector decomposition. Considering the diverse thresholds, the best results were obtained when only selecting the 10% of the total number of structural features. Regarding the integration alternatives, conversely to the results in the original paper and unlike for the *Twitter* dataset, the best results were obtained when integrating the utility matrices corresponding to each of the network dimensions. As it can be observed, *Tang et al.*'s results were lower than those of *Zalmout and Ghanem*, and hence lower that the results of our technique. Particularly, *Tang et al.*'s approach did not discover high quality communities. Interestingly, only one alternative achieved high *FlakeODF*, whilst none of them found semantically cohesive communities. As for *Zalmout and Ghanem*, discovered communities were not highly cohesive. As for our technique and *Zalmout and Ghanem* the best results were obtained when combining *Social* & *SimilarContent-0.6*. Nonetheless, despite considering the same relations, our technique was capable of finding communities of higher quality. In average, the quality differences ranged between 233% and 272% when considering the lowest and highest improvements for both graph derivations.

4.4.3. Summary of results

Fig. 17 compares the results obtained for the best performing node relationships combinations with the results of only considering the *Social* view. The depicted results are averaged across all the considered symmetrisation strategies. In most cases, consider-



(a) Independent Social and Content Views



(b) Weighted Social View

Fig. 16. *Flickr* dataset results – comparison to Tang et al. [33]'s approach.

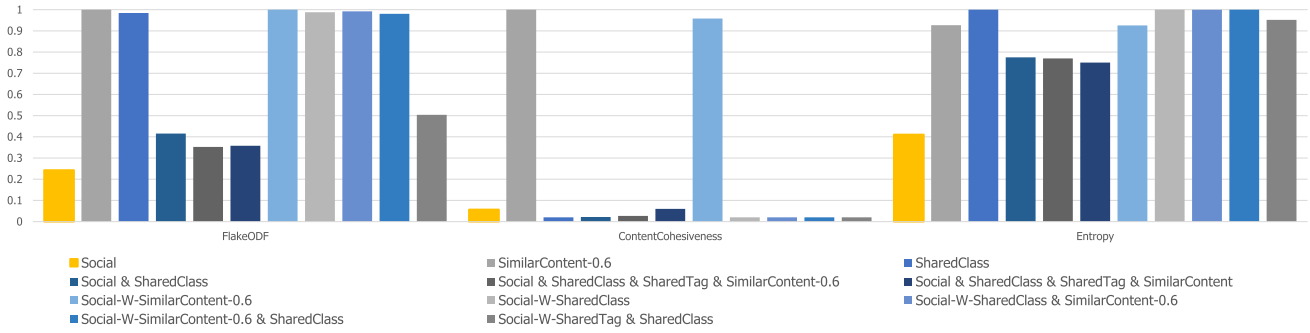
ing content improved results of only considering social information. Additionally, Table 7 summarises the improvements of the best performing relationship combinations over the *Social* view, the best *Zalmout and Ghanem* (*Social* & *SimilarContent-0.6* & *SharedClass* for the *Twitter* dataset and *Social* & *SimilarContent-0.6* for the *Flickr* one) and the best *Tang* et al. (*Social* & *SharedTag* & *SharedClass* & *SimilarContent-0.6* for the *Twitter* dataset, and *Social* & *SimilarContent-0.6* for the *Flickr* one) results. As it can be observed, the improvements over *Tang et al.* are higher than those over *Zalmout and Ghanem* for both datasets.

Regarding the *Twitter* dataset, the highest improvements with respect to only using the *Social* view, were obtained for both *ContentCohesiveness* and *FlakeODF*, which all combinations of relationships were able to outperform. These results imply that integrating content-based information to the community detection process always decreased the ratio of nodes that have more outer connections than inner ones. On the other hand, for the *Flickr* dataset, all combinations of relationships improved in average the quality of communities with respect to only using the *Social* view. As for the other dataset, the highest improvements were observed for the content cohesiveness of communities.

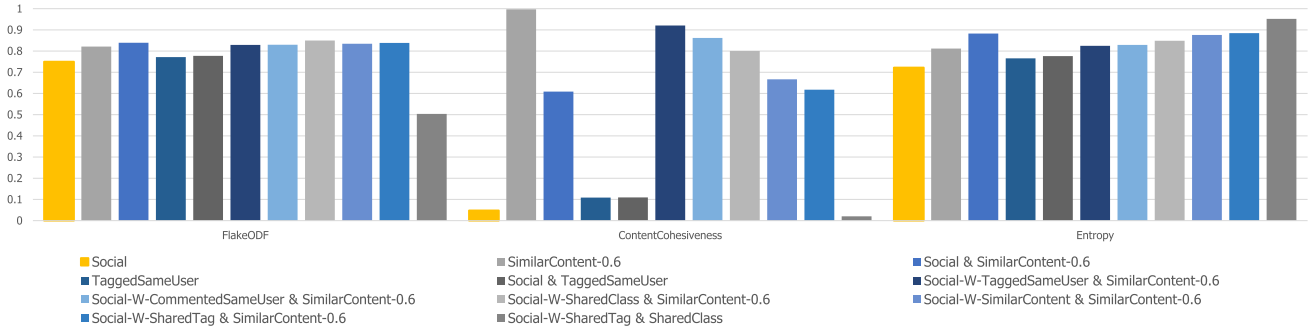
Considering the number of detected communities, most of the alternatives resulting in only one community were those combining all node relationships. These results agree with those in [32] that stated that considering multiple relations does not always improve quality results. This could be due to the fact that adding multiple relations creates a tightly connected and dense graph, which is difficult to partition. Conversely, in some cases individual relations led to an equal number of communities and nodes. It could be inferred that individual relations might not be sufficient to effectively partition graphs, as such relationships tended to create sparse graphs.

As regards the effect of the symmetrisation alternatives, results showed that the diverse strategies had a differentiated impact on the quality of the detected communities. Particularly, the Bipartite symmetrisation was shown to decrease the quality of communities regardless of the information sources under consideration for both datasets. On the other hand, the Simple and Bibliometric symmetrisations were shown to obtain similar results for the best performing combinations of relationships, showing that increasing the complexity of the symmetrisation alternative does not necessarily imply an improvement of the quality.

In summary, weighting the social information with the content-based relations achieved better results than independently combining them. Additionally, it is reinforced the necessity of adequately choosing not only which information sources to combine,



(a) *Twitter* Dataset



(b) *Flickr* Dataset

Fig. 17. Comparison of the best node relationship combinations.

Table 7
Summary of improvements (%).

	Flake-ODF			Content Cohesiveness			Entropy		
	Social	Zalmout and Ghanem	Tang et al.	Social	Zalmout and Ghanem	Tang et al.	Social	Zalmout and Ghanem	Tang et al.
(a) Twitter Dataset									
SimilarContent-0.6	311.89	79.03	69.61	1666.25	451.46	1299.80	125.77	19.41	157.89
SharedClass	440.57	79.03	69.61	−52.04	−89.37	−73.03	224.82	19.85	158.84
Social & SharedClass	71.13	−17.71	−22.04	−61.40	−76.76	−41.01	88.85	19.14	157.30
Social & SharedClass & SharedTag & SimilarContent−0.6	45.10	−20.75	−35.38	−52.55	−89.50	−73.34	87.58	19.07	156.86
Social & SharedClass & SharedTag & SimilarContent	47.48	−31.79	−24.92	5.88	−56.14	11.33	82.80	18.94	157.14
Social-W-SimilarContent-0.6	311.89	79.03	69.61	1592.31	428.38	1241.19	125.51	19.41	157.89
Social-W-SharedClass	442.58	79.03	69.61	−52.46	−56.25	11.06	224.82	19.85	158.84
Social-W-SharedClass & SimilarContent-0.6	444.88	79.03	69.61	−52.26	−56.25	11.06	224.68	19.85	158.84
Social-W-SimilarContent-0.6 & SharedClass	438.57	79.03	69.61	−52.32	−56.25	11.06	224.82	19.85	158.84
Social-W-SharedTag & SharedClass	176.61	−11.35	−16.01	−52.06	−56.25	11.06	209.17	19.75	158.61
(b) Flickr Dataset									
SimilarContent-0.6	9.49	0.60	7.07	1991.12	192.40	464.89	12.42	−0.35	120.30
TaggedSameUser	2.86	−1.53	4.80	127.63	−60.63	−23.94	6.04	−2.14	116.36
Social & SimilarContent-0.6	11.89	0.40	6.86	1176.77	53.52	196.58	22.22	10.54	144.39
Social & TaggedSameUser	3.64	−4.20	1.97	129.24	−62.15	−26.88	7.50	−0.55	119.86
Social-W-TaggedSameUser & SimilarContent-0.6	10.57	0.74	7.22	1831.25	165.30	412.55	14.20	1.79	125.03
Social-W-CommentedSameUser & SimilarContent-0.6	10.66	0.74	7.22	1707.55	138.07	359.94	14.81	2.82	127.33
Social-W-SharedClass & SimilarContent-0.6	13.32	0.08	6.52	1579.03	110.08	305.85	17.50	7.36	137.35
Social-W-SimilarContent & SimilarContent-0.6	11.27	0.31	6.76	1298.19	55.49	200.40	21.31	10.46	144.20
Social-W-SharedTag & SimilarContent-0.6	11.82	0.88	7.37	1195.88	53.51	196.58	22.51	10.43	144.13

but also how to combine them to effectively improve the quality of the detected communities.

The performed analysis over two real-world datasets allowed to show the benefits of applying the presented technique in comparison to state-of-the-art techniques, and to infer guidances for integrating multiple views for detecting communities in social networks. As regards the content-based relationships, results showed that a minimum similarity threshold should be imposed on content similarity for obtaining meaningful communities, as creating a full dense graph did not result in high quality communities. Additionally, the content of posts was reported to be more useful than the tags or hashtags assigned by users. Comments should be useful in those cases in which the goal is to find sentiment guided or polarised communities, otherwise they were shown not to be useful. Class or category information might be of interest, however, if no knowledge regarding whether the classes are determinant of the natural division of communities, they should not be used for linking the nodes. With respect to the topology-based relations, if the

degree of reciprocity of relations is unknown, applying a symmetrisation strategy is recommended. Adequately weighting the relations is important, as weighting them yielded better results than indiscriminately adding new information sources. It is worth noting that if the number of detected communities is close to the number of nodes in the graph (for example, in a graph of 100 nodes, each community has in average 2 or 3 nodes) or it is close to 1 (for example, finding 1 or 2 communities in a graph comprising 100 nodes) it is recommended to either remove some of the considered node relationships or modify their weights.

Finally, the intrinsic characteristics of the social network under analysis could also help to guide the selection of the relationships to consider. For example, on Information Oriented Networks, such as *Twitter*, content-based relations are more important than social relationships for finding high-quality communities. Nonetheless, social relations could also help to discover content-related communities. On the other hand, for Social Oriented Networks, in which the *Social* view is important, such as *Flickr*, only considering social relations might be sufficient for finding highly structurally connected communities. Furthermore, it is important to note that the friendship relationships is not the only source of topological information, the diverse social relations available should be explored in order to improve the quality of the found communities.

5. Conclusions

This work aimed at integrating multiple information sources for performing community detection in social networks. The proposed technique tackled the problem of how to combine several information sources for effectively finding high-quality community partitions. Moreover, it proposed several alternatives for adequately considering the semantics conveyed by directed relations.

Experimental evaluation conducted on two real-world social media datasets demonstrated that the different information sources offer complementary views of data. Each type of relation was shown to have a distinguished effect on the quality of the detected communities. Thus, results reinforced the fact that community detection techniques could benefit from the integration of multiple and diverse information sources. Furthermore, the strategies for conveying the semantics of directed relations also showed differentiated effects on community quality. However, results also showed that a naïve combination of information sources and symmetrisation strategies could result in low quality results, implying that the relations have to be carefully leveraged to achieve a positive effect on the quality of communities. Nonetheless, the study also showed that the diverse social networking sites have different motivations for the interactions between users (both social and content-based), which might affect the relevance of the information obtained through the different information sources. For example, *Twitter* was shown to be more content-driven, whereas *Flickr* showed a bias towards the underlying social relations. This implies that the intrinsic characteristics of social media data have to be taken into account when selecting the information to consider in the community selection process.

As regards future work, additional alternatives for considering the directionality of edges could be explored. For example, the metrics used for assessing the quality of the detected communities could be extended to consider edge directionality. Moreover, such metrics could be also extended to include a content cohesiveness assessment of communities. Regarding relation combination, the chosen graph representation collapses possibly heterogeneous information into a unique and homogeneous space, ignoring the possible differences amongst such relations. Hence, a multi-graph representation in which each relation is represented as a separated dimension could be devised. This representation would also allow optimising the community partition at each dimension individually. Additionally, it could be assessed whether it is beneficial to scale the weights of relations according to certain factors. Finally, the possibility of considering overlapping communities could be also studied.

References

[1] C.C. Aggarwal, K. Subbian, Event detection in social streams, in: Proceedings of the 2012 SIAM International Conference on Data Mining, 2012, pp. 624–635, doi:10.1137/1.9781611972825.54.
[2] V.D. Blondel, J. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. Theory Exper. 2008 (10) (2008) P10008.
[3] D.M. Boyd, N. Ellison, Social network sites: definition, history, and scholarship, J. Comput. Med. Commun. 13 (1) (2007) 210–230.
[4] R.A. Brualdi, F. Harary, Z. Miller, Bigraphs versus digraphs via matrices, J. Graph Theory 4 (1) (1980) 51–73.
[5] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, Nus-wide: a real-world web image database from national university of singapore, in: Proceedings of the ACM International Conference on Image and Video Retrieval, in: CIVR '09, ACM, New York, NY, USA, 2009, pp. 48:1–48:9.
[6] A. Clauset, M.E.J. Newman, C. Moore, Finding community structure in very large networks, Phys. Rev. E 70 (2004) 066111.
[7] P.M. Comar, P. Tan, A.K. Jain, A framework for joint community detection across multiple related networks, Neurocomputing 76 (1) (2012) 93–104. Seventh International Symposium on Neural Networks (ISNN 2010)Advances in Web Intelligence.
[8] G.W. Corder, D.I. Foreman, Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach, Wiley, New Jersey, 2009.
[9] P. Cui, W. Zhu, T.S. Chua, R. Jain, Social-sensed multimedia computing, IEEE MultiMedia 23 (1) (2016) 92–96.
[10] R.D. Silva, R. Stasiu, V.M. Orengo, C.A. Heuser, Measuring quality of similarity functions in approximate data matching, J. Inform. 1 (1) (2007) 35–46.
[11] H. Fang, F. Wu, Z. Zhao, X. Duan, Y. Zhuang, M. Ester, Community-based question answering via heterogeneous social network learning, in: D. Schuurmans, M.P. Wellman (Eds.), Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12–17, 2016, Phoenix, Arizona, USA., AAAI Press, 2016, pp. 122–128.
[12] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3–5) (2010) 75–174.
[13] M. Giatsoglou, D. Chatzakou, A. Vakali, Community detection in social media by leveraging interactions and intensities, in: WISE (2), in: LNCS, 8181, Springer, 2013, pp. 57–72.
[14] M.S. Granovetter, The strength of weak ties, Am. J. Sociol. 78 (6) (1973) 1360–1380.
[15] R. Guimerà, M. Sales-Pardo, L.A. Amaral, Module identification in bipartite and directed networks, Phys. Rev. E 76 (3 Pt 2) (2007) 036102.
[16] P.K. Reddy, M. Kitsuregawa, P. Sreekanth, S.S. Rao, A Graph Based Approach to Extract a Neighborhood Customer Community for Collaborative Filtering, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, pp. 188–200.
[17] J.G. Lee, P. Antoniadis, K. Salamatian, Faving reciprocity in content sharing communities: a comparative analysis of flickr and twitter, in: N. Memon, R. Alhajj (Eds.), ASONAM, IEEE Computer Society, 2010, pp. 136–143.
[18] 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, NYC, NY, USA, 2010, pp. 631–640.
[19] S. Lin, Q. Hu, G. Wang, P.S. Yu, Understanding Community Effects on Information Diffusion, Springer International Publishing, Cham, 2015, pp. 82–95.
[20] F.D. Malliaros, M. Vazirgiannis, Clustering and community detection in directed networks: a survey, CoRR (2013). Abs/1308.0971
[21] A. Marin, B. Wellman, Social network analysis: an introduction, in: J. Scott, P.J. Carrington (Eds.), The SAGE handbook of social network analysis, SAGE Publications Ltd., London, 2014, pp. 11–25, doi:10.4135/9781446294413.
[22] J. McAuley, J. Leskovec, Image Labeling on a Network: Using Social-Network Metadata for Image Classification, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 828–841.
[23] A. Mislove, H.S. Koppula, K.P. Gummadi, P. Druschel, B. Bhattacharjee, Growth of the flickr social network, in: Proceedings of the First Workshop on Online Social Networks, WOSN '08, ACM, New York, NY, USA, 2008, pp. 25–30.
[24] H.T. Nguyen, T.N. Dinh, T. Vu, Community detection in multiplex social networks, in: Proceedings of the 2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2015, pp. 654–659.
[25] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, A Graph-Based Clustering Scheme for Identifying Related Tags in Folksonomies, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 65–76.
[26] S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, A. Vakali, Cluster-based landmark and event detection for tagged photo collections, IEEE MultiMedia 18 (1) (2011) 52–63.
[27] Y. Pei, N. Chakraborty, K. Sycara, Nonnegative matrix tri-factorization with graph regularization for community detection in social networks, in: Proceedings of the 24th International Conference on Artificial Intelligence, AAAI Press, 2015, pp. 2083–2089.
[28] G.J. Qi, C.C. Aggarwal, T. Huang, Community detection with edge content in social media networks, in: Proceedings of the IEEE 28th International Conference on Data Engineering, 2012, pp. 534–545.

[29] V. Satuluri, S. Parthasarathy, Symmetrizations for clustering directed graphs, in: Proceedings of the EDBT, ACM, 2011, pp. 343–354.
[30] S.E. Schaeffer, Graph clustering, Comput. Sci. Rev. 1 (1) (2007) 27–64.
[31] J. Tang, H. Liu, Feature selection with linked data in social media, in: Proceedings of the 12th SIAM International Conference on Data Mining, SIAM / Omnipress, 2012, pp. 118–128.
[32] J. Tang, X. Wang, H. Liu, Integrating social media data for community detection, in: Proceedings of the LNAI, 7472, 2012, pp. 1–20.
[33] J. Tang, X. Wang, H. Liu, Modeling and mining ubiquitous social media: International workshops MSM 2011, boston, MA, USA, october 9, 2011, and MUSE 2011, athens, greece, september 5, 2011, revised selected papers, in: chapter Integrating Social Media Data for Community Detection, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 1–20.
[34] J.W. Tukey, Exploratory data analysis, in: Addison-Wesley Series in Behavioral Science, Addison-Wesley Publishing Company, 1977.
[35] X. Wang, D. Jin, X. Cao, L. Yang, W. Zhang, Semantic community identification in large attribute networks, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, in: AAAI'16, AAAI Press, 2016, pp. 265–271.
[36] K. Xu, K. Zou, Y. Huang, X. Yu, X. Zhang, Mining community and inferring friendship in mobile social networks, Neurocomputing 174 (Part B) (2016) 605–616.
[37] J. Yang, J. Leskovec, Defining and evaluating network communities based on ground-truth, Knowl. Inform. Syst. 42 (1) (2013) 181–213.
[38] T. Yang, R. Jin, Y. Chi, S. Zhu, Combining link and content for community detection: a discriminative approach, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, ACM, New York, NY, USA, 2009, pp. 927–936.
[39] J. Yang, J.J. McAuley, J. Leskovec, Community detection in networks with node attributes, in: Proceedings of the 2013 IEEE 13th International Conference on Data Mining, 2013, pp. 1151–1156. Dallas, TX, USA, December 7–10, 2013.
[40] N. Zalmout, M. Ghanem, Multidimensional community detection in twitter, in: Proceedings of the 8th International Conference for Internet Technology and Secured Transactions, 2013, pp. 83–88.
[41] Z. Zhang, Q. Li, D. Zeng, H. Gao, User community discovery from multi-relational networks, Decis. Support Syst. 54 (2) (2013) 870–879.
[42] F. Zhang, J. Li, F. Li, M. Xu, R. Xu, X. He, Community detection based on links and node features in social networks, in: Proceedings of the MultiMedia Modeling, Springer International Publishing, Cham, 2015, pp. 418–429.
[43] T. Zhang, P. Cui, C. Faloutsos, Y. Lu, H. Ye, W. Zhu, S. Yang, Come-and-go patterns of group evolution: a dynamic model, in: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, New York, NY, USA, 2016, pp. 1355–1364.
[44] Z. Zhao, L. Zhang, X. He, W. Ng, Expert finding for question answering via graph regularized matrix completion, IEEE Trans. Knowl. Data Eng. 27 (4) (2015) 993–1004.
[45] A. Zubiaga, D. Spina, R. Martínez, V. Fresno, Real-time classification of twitter trends, J. Assoc. Inform. Sci. Technol. 66 (3) (2015) 462–473.

Antonela Tommasel is a member of ISISTAN Research Institute (CONICET-UNICEN) since 2011. She received her Bachelor in Software Engineering at UNICEN University (Argentina) in November 2012 and a PhD in Computer Science degree at the same institution in December 2017. She is also a teacher assistant at the same university. Her research interests include recommender systems, text mining, user modelling and social web.

Daniela Godoy is a researcher at CONICET and a member of ISISTAN Research Institute, Tandil, Argentina. She is also a full-time professor in the Department of Computer Science at UNCPBA, Tandil, Argentina. She obtained her Master's degree in Systems Engineering (2001) and her PhD in Computer Science (2005) at the same university. Her research interests include intelligent agents, user profiling and text mining.