Consensus Community Detection for Multi-dimensional Networks

Antonela Tommasel*, Daniela Godoy†
ISISTAN, CONICET-UNICEN. Tandil, Buenos Aires, Argentina
*Email: antonela.tommasel@isistan.unicen.edu.ar
†Email: daniela.godoy@isistan.unicen.edu.ar

***Abstract*—Since their beginnings, social networks have affected the way people communicate and interact with each other. Nowadays, user interactions range from social relations to posting and reading activities, leading to the existence of multiple and complementary information sources or dimensions for characterising user behaviour. The task of community detection could benefit from integrating those multiple sources. However, most techniques disregard the effect of information aggregation, and continue to focus only on one aspect: network topology. This paper aims at providing some insights on how to integrate the multiple and heterogeneous social media information sources characterising user activities and behaviour to optimise the quality of found communities. To that end, diverse consensus strategies to extend techniques designed for a unique information source to multi-dimensional networks are presented and analysed. Experimental evaluation confirmed the benefits of using consensus strategies for leveraging on multiple data dimensions in terms of community quality.**

***Keywords*—Multi-dimensional Networks, Community Detection, Heterogeneous Information**

I. Introduction

Social networking and micro-blogging sites have increased their popularity in recent years attracting millions of users, who spend an increasingly amount of time on those sites sharing personal information and making new friends. For example, sites like *Flickr*, *YouTube*, *Facebook* or *Twitter* allow users to create content, publish photos, comment on content other users shared, tag content, and socially connect with other users in the form of subscriptions or friendships. Consequently, social networking sites affect how people communicate and interact. Unlike in the physical world, in social media people have a greater freedom to connect with a wider range of individuals for multiple reasons. In this context, social networks can be defined as a set of socially relevant nodes connected by one or more relations. Nodes are not only limited to representing people. Instead, they might also represent other entities such as social posts, tweets, geographical places or Web pages, amongst other possibilities.

The pervasive usage of social media offers research opportunities for analysing user behaviour and how users interact with their friends [17]. These research opportunities are not only related to computer science, but also to physics, economics, behavioural science and business marketing. One fundamental problem in social networks is identifying groups or communities of elements, when there is no explicit group information available [17]. Communities can be defined as sets of elements (i.e. users, posts or other elements) that interact more frequently or share more similarity with elements within their own community than with those outside of it. Community detection has been proved useful in diverse domains such as biology and social sciences, as they can be used for further analysis as visualisation, group profiling or relational learning.

Several techniques have been proposed to address the problem of detecting communities in networked data. However, most of them only focus on one individual aspect of users' relationships, whilst users interact or connect with others for many diverse reasons. For example, in *Twitter*, users might connect with others because they have the same interests, because they had retweeted or marked as favourite the same tweets, or because they share common friends. Additionally, users might post on the same topic or use the same hashtags. In this context, there are other sources or dimensions of information that might implicitly define connections between users in social media. Then, different networks can be built based on each form of information source or dimension, which combined result in a rich multi-dimensional network representing user activities and interactions.

When analysing a multi-dimensional network with heterogeneous information, one information source might be insufficient for accurately capturing community structure. For instance, in *Twitter*, social relations might be extremely sparse and two users might belong to the same community even if they do not follow each other. Relations can also be noisy since, as it is easier to connect with other users online than in the real physical world, users might have thousands of online friends. This situation could hinder the correct identification of the communities users belong to, if only friendship interactions are considered. On the other hand, other users might have a low number of friends, but frequently engage in other activities such as posting or commenting content, which could lead to valuable information for community detection. Nonetheless,

978-1-5386-3057-0/17/$31.00 ©2017 IEEE

social media content might be topically diverse and noisy for conveying valuable topic-based relationships. Thus, integrating multiple information sources could help to overcome the problems caused by incomplete and noisy information in each dimension, as well as obtaining more accurate and reliable community partitions. However, the combination of multiple and possible heterogeneous data dimensions (or views) poses new challenges. For example, how to fuse the different informational aspects provided by each dimension for performing an integrated analysis.

Considering the increasing amount of available information in social networks and the necessity of integrating such heterogeneous data, this paper addresses two challenges. First, the definition and extraction of multiple sources of information regarding user interactions and activities that can be inferred from social media data. Second, based on a multi-dimensional graph, this study presents and analyses four possible integration strategies for extending community detection techniques designed for a unique data dimension to leverage on multi-dimensional networks. The strategies are based on integrating the community structures found for each individual dimension into a single consensual community partition, reflecting the information provided by each dimension. The final goal of this study is to provide some insights on how to integrate the diverse information sources and user interactions for improving the quality of hidden community structures that are shared by the heterogeneous interactions.

The rest of this paper is organised as follows. Section II discusses related research. Section III defines the nature of the analysed dimensions in a multi-dimensional graph and presents the proposed consensus alternatives. Section IV describes the experimental settings and the obtained results. Finally, Section V summarises the conclusions drawn from this study and presents future lines of work.

II. Related Work

Social networks are usually represented by graphs comprising a set of nodes connected through links or edges. Such edges can be undirected (as friendships on *Facebook*) or directed (as the Followee/Follower relations on *Twitter* or *Instagram*). Communities can be regarded as potentially overlapping groups of nodes that are densely connected within the community, but sparsely connected with nodes outside of it. The goal of community detection techniques (also referred as graph clustering techniques) is to divide the nodes into groups (i.e. communities), such that the nodes in a community are similar or connected in a pre-defined sense [15]. Nonetheless, not every graph presents a natural community structure. For example, in a graph in which edges are evenly distributed over the set of nodes, the resulting clustering will be rather arbitrary.

Recently, the efforts of community detection techniques have been focused on addressing the challenges posed by the heterogeneous nature of social media by combining diverse social networks [12], or information sources, such as social and content information [21, 22, 19] or social, content and user similarity [14, 17]. Interestingly, in most approaches, all data dimensions are reduced to one dimension, hence merging the heterogeneous information sources into a unique graph, which can be problematic if dimensions are not comparable or have different relative importance. Only a few approaches [18, 3, 11] have leveraged on consensus techniques, which were not necessarily applied to multi-dimensional community detection.

Yang et al. [21] modelled the probabilities of nodes of being linked using conditional and content discriminative models for reducing the impact of irrelevant content features. Experimental evaluation was based on citation networks, where nodes corresponded to scientific articles, edges to citations and keywords described the content. Similarly, Zhang et al. [22] proposed a probabilistic model for combining topological information with node attributes. Experimental evaluation was based on *Twitter* and *Facebook* datasets from SNAP[1]. Hashtags and mentions were selected as the content features for *Twitter*, whilst the information in user profiles (e.g. hometown, birthday, political associations) was selected for *Facebook*. Optimisation was performed by means of Expectation Maximisation, which improved state-of-the-art techniques based on social links, content or a combination of both information sources.

Tommasel and Godoy [19] studied how to integrate multiple social and content-based information sources for discovering posts communities. In addition, the authors proposed alternatives for integrating edge directionality to the analysis. The considered content-based dimensions included tagging behaviour, posts' topics and posts' similarity. Content relations were used to weight the existing social relations, or to define other independent relations between posts. Results based on *Twitter* demonstrated that each information source offers complementary views, whose relevance depends on the characteristics of the networking site under analysis. Furthermore, results showed that naïvely combining information sources and edge semantics could lead to low quality results, implying that the relations have to be carefully leveraged to achieve a positive effect on community quality.

Tang et al. [17] discovered user communities by integrating multiple information sources in a joint optimisation problem. Social information was combined with the concatenation of all content-based sources. Evaluation was based on tags and comments from *BlogCatalog* and *Flickr*. Nodes represented users connected by friendship links. Results showed that the integration of multiple data sources introduced noise and redundancies, hence reducing the quality of communities, and increasing complexity.

Pei et al. [14] grouped users by combining topological information, content-based features, message similarity and user interactions in a non-negative matrix factor-

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

isation problem. Experimental evaluation was based on two small *Twitter* datasets comprising politicians. Unlike the previously presented works, results showed that social information performed better than content. Hence, the authors claimed that social relations are capable of accurately capturing user interests, whilst content information introduces noise. Nonetheless, as the evaluation was based on both social and content cohesive datasets, there is no guarantee that the assumptions would held on heterogeneous datasets where relations might respond to diverse and perhaps contradictory reasons.

Traditionally, consensus techniques are used for combining the results of several algorithms applied to the same network to improve individual results. In this context, Mathias et al. [11] proposed a genetic-based consensus algorithm for community detection in direct networks using modularity as the fitness function. The starting population of the algorithm is the communities found by diverse alternatives of Label Propagation. Then, populations are evolved until the modularisation criterion was met. Experimental evaluation based on the connections between republican and democrat blogs during 2004 showed that the approach was able to improve the results of Label Propagation and Infomap.

Burgess et al. [3] added information from missing edges to improve the quality of communities by combining a consensus clustering algorithm with link prediction. The proposed technique uses link prediction to build a probabilistic distribution over inferred edges. Then, it creates a set of networks from the built distribution, which are partitioned into communities by means of traditional community detection algorithms. Finally, the obtained partitions are aggregated into a network, in which the edge weights corresponded to the normalised frequency of the occasions in which two nodes belonged to the same community. The final communities are obtained by removing low confidence edges. Experimental evaluation based on a *Facebook* dataset from SNAP showed improvements over state-of-the-art techniques, at the expense of increasing the computational time.

Jin et al. [6] proposed a clustering fusion algorithm for detecting user communities in time evolving networks. To describe the time evolving characteristics of networks, the authors used snapshots to represent the network at a specific time. Nonetheless, the authors only considered the explicit social data for establishing users' relations. The algorithm comprises two steps. First, for each snapshot a base clustering is performed to obtain the clustering for that timestamps. Then, the obtained clusterings are merged to obtain the final clustering. Experimental evaluation based on a *Google+* dataset showed that although their algorithm obtained better results than the selected baselines, it was more time consuming.

Finally, closely related to this study and unlike previous works, Tang et al. [18] analysed four consensus strategies for discovering user communities in heterogeneous networks. First, all dimensions were integrated into one by averaging the individual weights. Second, the objective function was simultaneously optimised over all dimensions. Third, features from each dimension were extracted and then PCA was applied to capture the principal patterns across all dimensions. Fourth, multiple clustering results were combined into a single consensual clustering. Experimental evaluation was based on a *YouTube* dataset that included five data dimensions: the users' social relations, friends-of-friends, co-subscription, co-subscribed and cofavourite. Results showed that independently considering each data dimension achieved better results than collapsing all dimensions into a unique graph, but worse results than the other integration strategies.

III. Consensus Community Detection based on Heterogeneous Social Information

To apply a community detection algorithm, the information on which the underlying graph structure is going to be based on has to be defined. Multiple and diverse information sources can be extracted from social media data, and hence multiple graph structures can be defined. Nodes might not only represent real people, but also other entities such as neighbourhoods, Web pages or tweets, amongst others depending on the task to perform [10]. For example, if the goal of finding communities is to predict new relations between users or to measure the influence users have on their neighbourhood [20], the nodes in the graph might represent actual users of the network. Conversely, if the goal is to discover relations between tags in a folksonomy [13], nodes could represent tags. The goal of this work is to detect communities of related posts in social media, hence, as in [1, 21], each node in the graph represents a social post. Once communities are found, they can be integrated in diverse learning tasks such as topic detection, text classification or clustering, link prediction, or even feature selection.

In social media networks, users can not only establish social relations with other users, but also, create content. Consequently, social media data can be regarded as an heterogeneous network comprising not only information in the form of social (i.e. friendship or followee/follower) relationships, but also other information sources representing other types of relations between users or posts. For example, showing interest in a post by bookmarking it or the exchange of comments and tags could denote other sources of relationships between users. These activities provide different points of view of the same network, hence being useful for finding the community structure of a network. Nonetheless, such different types of relations need to be adequately leveraged when representing them in the graph. In this context, Section III-A presents the diverse node relations that could be considered when creating the social graph representation of the network. Then, Section III-B defines how to represent the social graph structure once all node relations are defined.

A. Graph Extraction

Most community detection techniques are purely based on the social relations amongst the elements in the underlying social media network. However, in the context of social media data, both the social relations between users and the characteristics of the published content are important for improving the quality of the discovered communities. Hence, besides the relations between posts derived from the actual social relations between their authors (i.e. two posts are socially related if their are authors are socially connected), posts' content resemblance or common categories (in case they are available) could also help to establish relations between them.

Additionally, the specific characteristics and metadata from each micro-blogging site could be exploited for discovering other meaningful relations between posts. For example, the usage of hashtags is encouraged in *Twitter*, *Instagram* and *Facebook* to aid in the search of messages of a specific theme or content. Then, posts containing the same (or associated) hashtags could be assumed to be topically related. As explained, social and content-based relations offer complementary data views, thus, no individual relation alone might be sufficient for determining high quality community partitions [17]. For instance, social information might be sparse and noisy, whilst content information might be irrelevant or redundant.

Content-based relations could be used either to reinforce the social relations already found amongst posts or to establish new relations amongst posts that are not socially connected. In the former case (weighted graph derivation), the graph only includes edges representing the social relations between nodes, whose relevance is given by the content features. Consequently, the quality of the social ties between nodes depends on the adequate definition of the content features in order to fully exploit all information sources. In the latter case (independent graph derivation), social and content relations are assumed to be independent and hence, edges in the graph represent either social or content links. As a result, two nodes might be connected even if there is no explicit social connection between them. For the purpose of this work, several content-based relations are defined based on the information available on social networking sites:

- *Shared Tags.* An edge between two nodes exists if they share any tag (or hashtag). The weight of the edge is measured as the 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 weight of edges could be computed as the distance between both categories.
- *Similar Content.* Measures the content resemblance of two nodes. A minimum similarity threshold might 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, it could be expressed by simply computing the percentage of shared terms between the two nodes, or by computing their Cosine Similarity.

By definition, all content-based relations are symmetric, i.e. they do not have directionality. However, regarding the Follower/Followee relationship in *Twitter*, the fact that user A follows user B does not imply that user B also follows user A. The same applies to the social relations in *Instagram*. Whilst the diverse social networks exhibit different reciprocity levels, most community detection techniques only leverage on undirected (and perhaps weighted) graphs. It is worth noting that developing community detection techniques for directed graphs might be a difficult task [5], and that several concepts that are theoretically well defined for undirected graphs have not been yet extended to directed ones [9]. Hence, for the purpose of this work, the directionality of social relations was ignored, transforming the directed graph in an undirected one in order to employ techniques already defined for undirected graphs. This applied transformation is one of the simplest and most commonly symmetrisation techniques.

Once all relations amongst nodes are defined, they can be represented in a simple (or unique) graph or a multigraph. The simple graph collapses multiple (and possibly heterogeneous) relations between two nodes into a unique edge, i.e. if multiple relations exists between two nodes, such relations are collapsed into a unique edge. This alternative ignores the differences amongst heterogeneous spaces. On the other hand, in a multi-graph, each relation between nodes is represented as a separated dimension of the same graph. This last representation allows to treat dimensions separately, allowing to individually optimise the community structure of each particular dimension. For the purpose of this work, the multi-graph representation was selected.

B. Finding Communities in a Multi-dimensional Graph

Traditional community detection techniques are designed for assessing only one dimension of the graph at the time, or the simple and collapsed representation of the graph dimensions. However, as previously mentioned, such representation collapses possibly heterogeneous information into a unique and homogeneous space, ignoring the differences amongst such dimensions. An alternative to cope with this limitation is to consider each graph dimension as a separated graph and discovering the community distribution for each of them separately. Then, the obtained results could be aggregated by means of a cluster ensemble strategy. Cluster ensemble or consensus clustering refers to the situation in which multiple clusterings, or

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

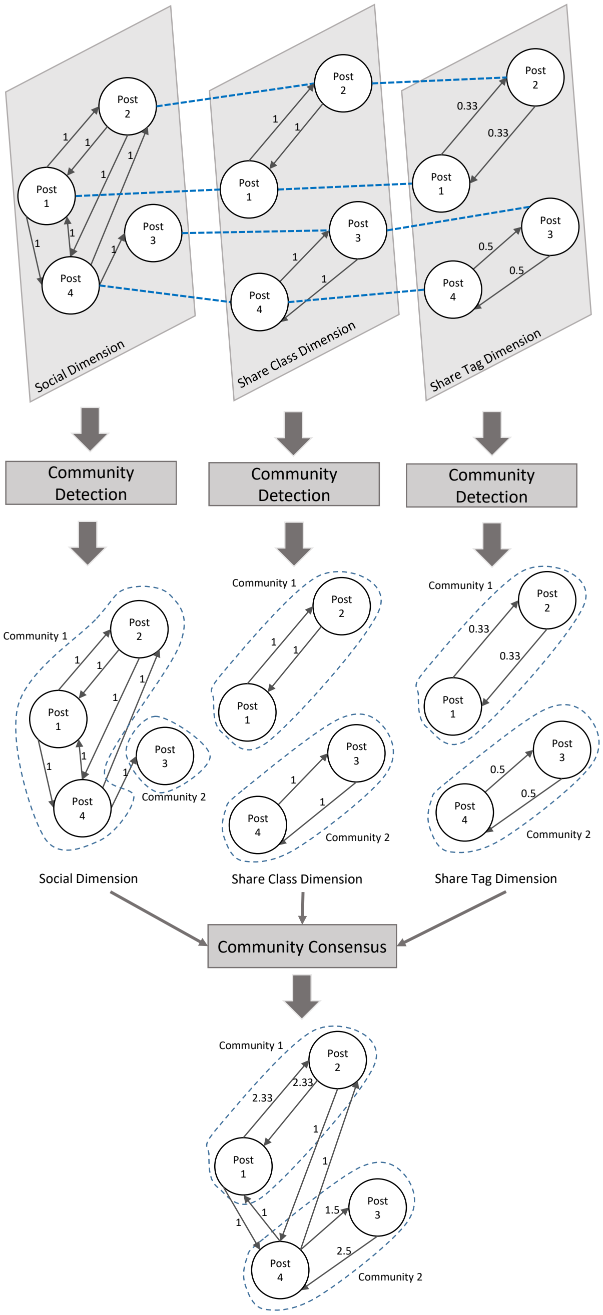


Figure 1: Consensus Community Detection

community partitions, have been obtained and it is desired to find a single clustering or community partition, which is a better representative of the individual community structures [7]. Figure 1 depicts an example of applying consensus techniques to the communities obtained for each individual dimension of the defined multi-graph in order to obtain a final community partition that integrates the information of all information sources.

Several alternatives have been proposed in the literature for combining a given ensemble of community partitions to produce the final community partition [16, 4], which have not yet been applied to the context of social media data. This work analyses four consensus alternatives: instance-based, cluster-based, hybrid-bipartite and metric-based. The focus of the instance-based, cluster-based and hybrid-bipartite alternatives is on building an intermediate representation comprising all the partitions found, and then applying once again a clustering or community detection technique over the newly created graph. In general, the alternatives do not require access to the original graph features nor the technique that determined such partitions. In all cases, once the intermediate representation is built, it is fed to the selected community detection algorithm for obtaining the final community partition.

Instance-based Consensus: The instance-based alternative [16] builds a graph that models the pairwise similarity amongst the original data instances. Each data instance is represented as a node in the new graph. Information regarding the similarity amongst communities is ignored. The edges between nodes are weighted proportionally to how frequently the two nodes are located in the same community. Particularly, three alternatives for weighting the edges are proposed. First, the number of shared communities (named *instance-based*). Second, the percentage of shared communities amongst the total number of communities (named *instance-based-%*). Third, the number of shared communities scaled by the nodes' content similarity (named *instance-based-content*). Note that whilst the first and second alternatives are only based on the structure of the community partitions, the third alternative includes information regarding the content of nodes. Once the graph is built, the partition resulting from that graph can be regarded as the final community partition.

Cluster-based Consensus: The cluster-based alternative [16] builds a graph that models the correspondence, i.e. the similarity, amongst different communities in a given ensemble. Each community in each partition is represented as a node in the new graph. The edges between nodes are weighted according to the Jaccard Similarity, which is computed considering the data instances in each community. Once that newly created graph is partitioned, the final community partition is obtained as follows. First, each group of communities represents a meta-community. Each original data instance is assumed to be associated to a meta-community if such meta-community includes a community the data instance belongs to. Note that an instance might be associated with several meta-communities. In such cases, an instance is assigned to the meta-community to which it is most frequently associated. This strategy assumes that there exists a structural correspondence amongst the different community partitions found in the ensemble [4], i.e. the different partitions show some degree of similarity. This assumption might affect the quality of the resulting communities in those cases in which there is no such structural correspondence.

Hybrid-bipartite Consensus: The hybrid-bipartite alternative [4] creates a bipartite graph, in which nodes represent both the original data instances and the com-

Table I: *Twitter* Data Collection Main Characteristics

Number of Instances	1,036
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

munities. As this graph is intended to be bipartite, edges only connect nodes representing data instances with nodes representing communities, according to whether the data instance is included in the community. All edges have weight 1. This alternative simultaneously considers the instance and community similarity when obtaining the final community distribution.

Metric-based Consensus: The metric-based alternative does not actually combine the different partitions. Instead, it assesses the quality of the different partitions in terms of a quality metric. Then, the partition achieving the highest quality results is selected as the final partition. Particularly, quality is assessed by two metrics. First, betweenness centrality (named *metric-based-betweenness*). Second, the average content similarity of communities (named *metric-based-content*).

IV. Experimental Evaluation

This section presents the experimental evaluation performed to assess the effectiveness of the proposed approach for finding communities in heterogeneous social media data, and is organised as follows. Section IV-A presents the data collection used. Section IV-B presents implementation details, and the metrics used for evaluating the effectiveness of the different alternatives. Finally, Section IV-C presents the results derived from the performed experimental evaluation.

A. Data Collection

The performance of the technique was evaluated considering a real-world dataset collected from *Twitter*[3] [23]. It included the content of more than $500,000$ tweets belonging to $1,036$ trending topics, which were manually assigned to one of four categories: News, Ongoing Events, Memes (trending topics 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 I summarises the main characteristics of the dataset. For the purpose of the experimental evaluation, each trending topic was regarded as a node in the graph, i.e. each node grouped the tweet set associated to the corresponding trending topic.

B. Experimental Settings

The Java programming language was chosen for implementing the approach. Communities were found by the Gephi[4] implementation of the Louvain algorithm [2].

Table II: Evaluated Combinations of *Social* and *Content-based* Relationships

Independent Graph Derivation
Social & *SimilarContent-0.6*
Social & *SharedClass*
Social & *SharedClass* & *SharedTag* & *SimilarContent-0.6*
Social & *SharedClass* & *SharedTag* & *SimilarContent*

Weighted Graph Derivation
Social-W-SimilarContent-0.6 & *SharedClass*
Social-W-SharedClass & *SimilarContent-0.6*
Social-W-SharedTag & *SharedClass*

Nonetheless, both the graph representation model and consensus strategies can be used in combination with any other community detection algorithm or technique. Evaluation was performed based on the diverse combinations of the social and content-based relations presented in Section III. A social relation (named *Social*) between two nodes was established if the authors of the tweets of a node followed authors of the other node. Additionally, two variations of the *SimilarContent* relation were defined: 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*). The chosen combinations correspond to the ones obtaining the best results in terms of a collapsed graph representation [19]. The selected combinations of social and content-based relations were evaluated in the context of two experimental settings. First, a setting in which content and social relations are independently considered, i.e. relationships between nodes can represent either social or content-based relations. Second, a setting in which only social relations between posts are established, and the content-based information is used to determine the importance of such relations. Table II summarises the evaluated combinations of relationships corresponding to either the independent or weighted graph derivations. The results of combining the multi-graph representation with consensus strategies for accurately finding community partitions are compared to those obtained for a collapsed representation of the social graph.

To assess 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 $1,000$ posts) were considered in the experimental evaluation. For each graph size, five random partitions were generated.

The quality of the discovered community partitions was evaluated by three scoring functions. First, as communities are built on the assumption that they comprise sets of nodes with many inner connections and few outer connections, *Flake-ODF* (Out Degree Fraction) Leskovec et al. [8], which is a function characterising community connectivity structure. Second, a function characterising the content cohesiveness of communities, i.e. the average Cosine Similarity between all node pairs in the com-

[3] http://www.twitter.com/

[4] http://gephi.github.io/

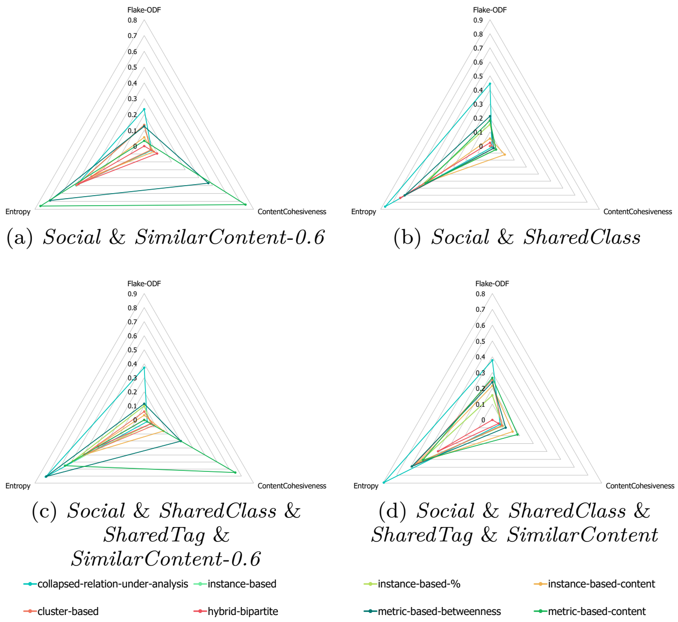
munity (named *ContentCohesiveness*). Third, as the selected dataset includes class assignments for each post (i.e. the class assigned to each trending topic), the *Entropy* of classes given the community assignments. As scores were individually computed for each discovered community, they were averaged to obtain the score corresponding to a given community partition. To ensure metrics' comparability, all results were normalised to the range $[0;1]$, and adjusted so that the highest scores represent the best ones.

C. Experimental Results

Considering the results obtained for each of the evaluated graph sizes, it was analysed whether the proposed relationships and symmetrisation strategies behaved stable across such sizes. As data did not follow a normal distribution, the Kruskall-Wallis test for unrelated samples was applied to each metric's results. The confidence value was set to 0.05. To perform the tests, the null and the alternative hypotheses were defined. The null hypothesis stated that no difference existed amongst the results of the different samples, i.e. the alternatives behaved stable over the different graph sizes. On the contrary, the alternative hypothesis stated that changes in the graph size caused changes in the behaviour of the alternatives. As in all cases the p-value was higher than the confidence value, the null hypothesis could not be rejected. Thus, the results obtained for the different graph sizes can be summarised by their mean values.

Figures 2 and 3 present the obtained results grouped by the graph derivation analysed and the selected combinations of relationships. In the figures, *collapsed-relation-under-analysis* represents the results obtained for the corresponding combination of relationships when considering a collapsed graph representation. It is worth noting that for some combinations, not every consensus alternative found a meaningful number of communities, i.e. a number between 1 and the total number of analysed nodes. Hence, those results are not reported. For every combination of relationships tested, the diverse consensus strategies have differentiated effects over the evaluation metrics. Section IV-C1 presents the results for the independent graph derivation, whilst Section IV-C2 for the weighted graph derivation.

1) Independent Social and Content Views: When considering the combination of *Social* & *SimilarContent-0.6* (Figure 2a) none of the consensus alternatives was able to improve the *Flake-ODF* results of considering the collapsed graph representation, meaning that the found communities had a higher fraction of nodes having more edges towards nodes outside the community than to nodes inside it. In other words, the found communities were not highly separated. Nonetheless, in all cases, the consensus strategies improved the quality of communities regarding *Entropy* and *ContentCohesiveness*. As a result, finding the independent community structure for each of the relations and then combining them, allowed to adequately leverage on the content-based relations to increase the content relatedness of posts inside the communities.



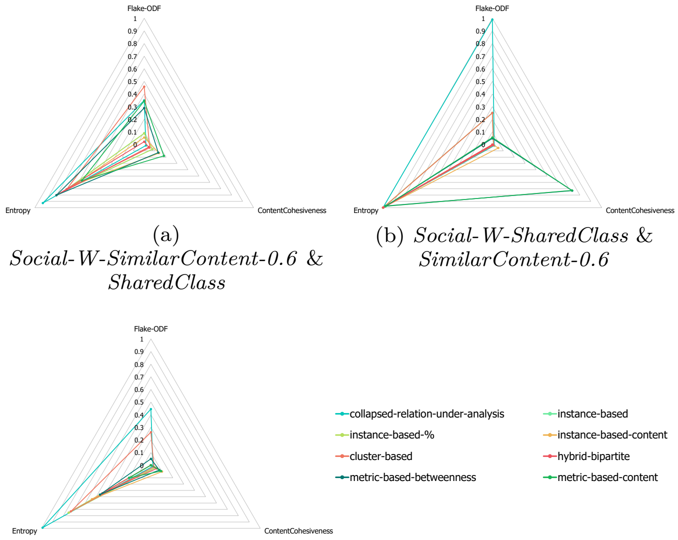
(a) *Social* & *SimilarContent-0.6*

(b) *Social* & *SharedClass*

(c) *Social* & *SharedClass* & *SharedTag* & *SimilarContent-0.6*

(d) *Social* & *SharedClass* & *SharedTag* & *SimilarContent*

Figure 2: Evaluation Results - Independent Social and Content Views



(a) *Social-W-SimilarContent-0.6* & *SharedClass*

(b) *Social-W-SharedClass* & *SimilarContent-0.6*

(c) *Social-W-SharedTag* & *SharedClass*

Figure 3: Evaluation Results - Weighted Social View

Interestingly, *metric-based-betweenness* achieved better *ContentCohesiveness* results than the other alternatives (with the exception of *metric-based-content*), showing that it is not necessary to explicitly assess the content relatedness of communities to find content cohesive communities. The *cluster-based* and *instance-based* strategies weighting the edges according to the absolute and average number of shared communities obtained the same quality results, which worsened when scaling those weights according to the content similarity of nodes. In this case, the content quality of communities was not improved regarding that of considering the collapsed graph. Finally, *hybrid-bipartite* obtained the worst results, showing that it is not only important to adequately choose the node relations to consider, but also to adequately combine the found community structures to optimise community quality.

As regards the combination of *Social* & *SharedClass* (Figure 2b), the consensus alternatives were not able to improve the community quality obtained for the collapsed graph representation, regarding neither *Flake-ODF* nor *Entropy*. However, they were able to improve *Content-Cohesiveness*. Similarly to the previous case, the best overall results were obtained with the *metric-based* consensus alternatives, even when failing to find the most content cohesive communities. The results obtained for this particular combination of relations reinforced the complementary nature of the diverse information sources. As applying consensus strategies to communities found by independent relations achieved worse results than their collapsed graph representation, the information provided by each independent relation might not be enough for uncovering the community structure. Considering the nature of social relations between users, this could mean that users socially interact with others who do not necessarily posts regarding the same topics, and that not every pair of users posting regarding a particular topic is socially related. In turn, these individual relations could lead to sparse and noisy graphs, thus hindering the accurate discovery of communities, as stated in [17]. This is highlighted by the fact that *cluster-based* consensus was unable to find a meaningful number of communities, highlighting the difficulty of finding communities for this particular combination of relations.

Even when assessing the content information of communities in the consensus strategies, their final *Content-Cohesiveness* was lower than when explicitly considering content as a relation. Nonetheless, in both cases, *Entropy* was relatively high. This could imply 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, Commemoratives, Memes and Ongoing Events) that do not actually represent posts' topics, i.e, two post could belong to the same category but contain unrelated content.

As it can be observed in Figure 2d, for *Social* & *SharedClass* & *SharedTag* & *SimilarContent-0.6*, results are similar as those obtained for *Social* & *SimilarContent-0.6* (Figure 2a). In both cases the *metric-based* strategies found highest quality communities, which improved the *Entropy* and *ContentCohesiveness* of the communities found with the collapsed graph. Similarly, *hybrid-bipartite* consensus obtained the worst results.

When comparing the results obtained with the collapsed representations of *Social* & *SimilarContent-0.6* (Figure 2a) and *Social* & *SharedClass* & *SharedTag* & *SimilarContent-0.6* (Figure 2d), it is observed that the latter allowed to find communities with better *Flake-ODF* and *Entropy*, whilst maintaining their *ContentCohesiveness*. In other words, adding more relations to the collapsed graph allowed to improve communities' quality. However, when comparing the results obtained for the different consensus strategies it is observed that those based on the former combination allowed to obtained communities of higher quality than when based on the latter. This could respond to two different situations. First, the consensus alternatives were unable to leverage on all the information provided by every analysed relation. Second, the individual relations provided noisy information that increased the difficulty of the community detection process. In any case, these results emphasise the importance of adequately choosing the relations to consider, and how choosing noisy, redundant or even contradictory relations could affect the outcome of the techniques.

Even though *Social* & *SharedClass* & *SharedTag* & *SimilarContent* (Figure 2c) only differs from the previously analysed relation in the threshold imposed for the *SimilarContent* relation, its results were more similar to those obtained for *Social* & *SharedClass* (Figure 2b) than those of *Social* & *SharedClass* & *SharedTag* & *SimilarContent-0.6* (Figure 2d). The tendency of the consensus strategies was the same as for the other relations as the best results were obtained with the *metric-based* alternatives and the worst ones with the *hybrid-bipartite* consensus.

Although including *SimilarContent* improved the *ContentCohesiveness* of communities, the improvements were lower than for *SimilarContent-0.6*. These results reinforce the fact that content-based relations could also introduce noise if not carefully analysed, and hence highlighted the importance of imposing a minimum threshold of similarity for regarding two nodes as content-related. However, changing the threshold of *SimilarContent* allowed to find communities with higher *Flake-ODF* than when considering *Social* & *SharedClass* & *SharedTag* & *SimilarContent-0.6* (Figure 2d). When comparing the results to those of *Social* & *SharedClass* (Figure 2b), the addition of more relations to analyse to the consensus strategies did not lead to better *Entropy* results. This situation might imply that tag and indiscriminate content information can misguide the algorithm in finding communities of posts belonging to the same category.

2) Weighted Social View: Figure 3a depicts results for the combination of *Social-W-SimilarContent-0.6* & *SharedClass*. As the Figure shows, the *cluster-based* consensus strategy improved the *Flake-ODF* of communities. With respect to the *ContentCohesiveness* obtained for the collapsed representation, its results were improved by every consensus strategy. However, none of the consensus alternatives was able to improve the *Entropy* of communities. Similarly to the previous cases, *metric-based* consensus obtained the best overall results.

The results of the combination *Social-W-SharedClass* & *SimilarContent-0.6* (Figure 3b) are the most representative of the importance of choosing both the adequate node relations, and whether to consider the collapsed graph representation or use consensus to combine the diverse relationships. Both the collapsed graph representation and the diverse consensus strategies achieved maximum *Entropy*. On the contrary, most of the consensus strategies obtained close to zero results for *Flake-ODF*, whilst the collapsed representation obtained the optimal ones. Note that, as in the previously analysed combinations of relations, only *metric-based* consensus was able to find high content cohesive communities.

Finally, Figure 3c presents the results obtained for *Social-W-SharedTag* & *SharedClass*. Although the collapsed relations lead to similar results than the other analysed relations regarding both *Flake-ODF* and *Entropy*, the consensus strategies obtained the worst results in comparison to those obtained for the other combinations of relationships. These results might indicate the fact that these specific individual relations do not provide enough information for the community detection algorithm, hence leading to the discovery of low quality communities.

From the results of *Social-W-SharedClass* & *SimilarContent-0.6* (Figure 3b) and *Social-W-SharedTag* & *SharedClass* (Figure 3c), it can be inferred than when integrated into the same graph, the information provided by the diverse relations helps to create a cohesive and complementary graph by strengthening or creating new links between nodes. On the other hand, when analysing each relation individually, the nature of each of them might lead to sparse graphs or even completely different graph structures that might mislead the consensus algorithm, hence resulting in low quality communities.

D. Summary of Results

From the performed analysis of the results obtained for each of the diverse combinations of relationships it can be inferred the effect that choosing the wrong consensus strategy can have over the characteristics of the discovered communities. Particularly, in all cases the highest quality communities were obtained when using *metric-based* consensus strategies, whilst *hybrid-bipartite* consensus obtained the worst community partitions even when including information regarding both the individual instances and the found communities.

It is worth noting that using consensus strategies to combine the communities found by individual node relations did not always yield the highest quality results. Results showed that for some combinations of relations, the highest quality communities were obtained when collapsing relations into a unique graph. In those cases, the consensus strategies allowed to obtain communities of the same or worse quality. This situation could respond to different reasons.

First, the information provided by each individual relation might not be neither enough nor accurate for discovering the community structure. As stated in [17], relations could be noisy or sparse, hindering the community detection process. For example, the *SharedTag* and *SimilarContent* might result in similar graph structures, which leads to similar community structures that do not add any extra information to the process. Second, relations might be redundant. In such cases, adding more relations does not necessarily imply adding new information to the process. Third, relations might provide contradictory information regarding the nodes. For example, the *SimilarContent* and *SharedClass* relations. For the analysed dataset, the topic or class assigned to posts does not have a correlation towards the actual content of posts. In this regard, whilst two posts might be linked by the *SimilarContent* relation, they might not be linked by the *SharedClass* one.

In summary, the selection of the relations to consider and whether to collapse all relations into a unique graph or apply a consensus strategy should be guided by the characteristics of the data under analysis. For example, as *Twitter* is a social platform aimed at sharing information, content-based relations might convey more information than the social ones, which might be rather casual or noisy. Additionally, as results showed, in those cases in which more than two relations are meant to be used, it is likely that the same results would be obtained with fewer relations (thus reducing the computational complexity) given the selection of the adequate consensus strategy.

In those cases in which relations might be contradictory (as in the case of the *SimilarContent* and *SharedClass*), it might be preferable to collapse them into a unique graph, as the nature of relations might mislead the consensus strategy. For example, *cluster-based* and *instance-based* techniques were shown to be highly sensitive to differences on the underlying community structures. Thereby, community quality tended to be low. Finally, in the overall case, *metric-based* techniques were shown to attain the best results. As those techniques rely on computing a score for each community structure they are not susceptible to the existence of redundant information (redundant structures would achieve similar scores) nor to the existence of contradictory information (the computed metric would favourite one structure over the other) in the general case.

V. Conclusions

Based on the increasing amount of information available in social networks, which leads to the necessity of integrating heterogeneous relations between users and posts, this work aimed at analysing several consensus strategies for extending community detection techniques designed for a unique data dimension to multi-dimensional networks. It tackled the problem of how to combine diverse information sources available in social media data to optimise the quality of the discovered communities. Moreover, it showed the effect that each defined consensus strategy has over community quality, and how the information sources behave when individually considered.

Experimental evaluation showed interesting findings. First, using consensus strategies could help to improve the quality of communities with respect to collapsing multiple heterogeneous relations into a unique graph. Second, the diverse information sources were found not to evenly contribute to the improvement of community quality, showing that including multiple information sources does not necessarily imply performance improvements. Third, the diverse consensus strategies have a distinct effect over the quality of communities. Fourth, the behaviour of the information sources differs according to whether they are mixed together or individually analysed. Thereby, the selection of the diverse information sources should be in accordance to how they are going to be integrated, or used, in the community detection process.

Finally, some interesting problems for further exploration could be derived from the performed analysis. First, the effect of the consensus strategies over other set of node relationships could be studied. Second, the nature and characteristics of each set of information sources could be studied to determine whether it is convenient to apply consensus techniques or collapsing them into a unique graph. Third, the performance of applying more advanced consensus strategies could be assessed.

References

[1] C.C. Aggarwal and K. Subbian. *Event Detection in Social Streams*, chapter 53, pages 624–635. 2012.

[2] V.D. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[3] M. Burgess, E. Adar, and M. Cafarella. Link-prediction enhanced consensus clustering for complex networks. *PLoS ONE*, 11(5):1–23, 05 2016.

[4] X.Z. Fern and C.E. Brodley. Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of the 21st International Conference on Machine Learning*, ICML '04, NY, USA, 2004. ACM.

[5] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.

[6] R. Jin, C. Kou, and R. Liu. Improving community detection in time-evolving networks through clustering fusion. *Cybernetics and Information Technologies*, 15(2):63–74, 2015.

[7] A. Lancichinetti and S. Fortunato. Consensus clustering in complex networks. *Sci. Rep.*, 2, 2012.

[8] J. Leskovec, K.J. Lang, and M. Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 631–640, New York, NY, USA, 2010. ACM.

[9] F. D. Malliaros and M. Vazirgiannis. Clustering and community detection in directed networks: A survey. *CoRR*, abs/1308.0971, 2013.

[10] A. Marin and B. Wellman. *Social Network Analysis: An Introduction*. Sage, London, June 2009.

[11] S.B. Mathias, V. Rosset, and M.C.V. Nascimento. Community detection by consensus genetic-based algorithm for directed networks. *Procedia Computer Science*, 96:90 – 99, 2016.

[12] H.T. Nguyen, T.N. Dinh, and T. Vu. Community detection in multiplex social networks. In *Computer Communications Workshops (INFOCOM WKSHPS), 2015 IEEE Conference on*, pages 654–659, April 2015.

[13] S. Papadopoulos, Y. Kompatsiaris, and A. Vakali. *A Graph-Based Clustering Scheme for Identifying Related Tags in Folksonomies*, pages 65–76. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[14] Y. Pei, N. Chakraborty, and 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*, pages 2083–2089. AAAI Press, 2015.

[15] S.E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27 – 64, 2007.

[16] A. Strehl and J. Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617, 2003.

[17] J. Tang, X. Wang, and H. Liu. *Integrating social media data for community detection*, volume 7472 of *Lecture Notes in Computer Science*, pages 1–20. 2012.

[18] J. Tang, X. Wang, and H. Liu. *Modeling and Mining Ubiquitous Social Media: International Workshops MSM 2011*, chapter Integrating Social Media Data for Community Detection, pages 1–20. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[19] A. Tommasel and D. Godoy. Integrating heterogeneous information from social networks into community detection. In *4th IJCAI Workshop on Heterogeneous Information Network Analysis (HINA)*, New York, NY, USA, 2016.

[20] K. Xu, K. Zou, Y. Huang, X. Yu, and X. Zhang. Mining community and inferring friendship in mobile social networks. *Neurocomputing*, 174:605 – 616, 2016.

[21] T. Yang, R. Jin, Y. Chi, and 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, pages 927–936, New York, NY, USA, 2009. ACM.

[22] F. Zhang, J. Li, F. Li, M. Xu, R. Xu, and X. He. Community detection based on links and node features in social networks. In *MultiMedia Modeling*, pages 418–429, Cham, 2015. Springer International Publishing.

[23] A. Zubiaga, D. Spina, R. Martínez, and V. Fresno. Real-time classification of twitter trends. *Journal of the Association for Information Science and Technology*, 66(3):462–473, 2015.