This paper was presented as part of the Workshop on Network Science for Communication Networks (NetSciCom)

# *ISCoDe*: a framework for interest similarity-based community detection in social networks

Eva Jaho, Merkouris Karaliopoulos and Ioannis Stavrakakis
Department of Informatics and Telecommunications
National & Kapodistrian University of Athens
Ilissia, 157 84 Athens, Greece
Email: {ejaho, mkaralio, ioannis}@di.uoa.gr

*Abstract*—**This paper proposes a framework for node clustering in computerized social networks according to common interests. Communities in such networks are mainly formed by user selection, which may be based on various factors such as acquaintance, social status, educational background. However, such selection may result in groups that have a low degree of similarity. The proposed framework could improve the effectiveness of these social networks by constructing clusters of nodes with higher interest similarity, and thus maximize the benefit that users extract from their participation. The framework is based on methods for detecting communities over weighted graphs, where graph edge weights are defined based on measures of similarity between nodes' interests in certain thematic areas. The capacity of these measures to enhance the sensitivity and resolution of community detection is evaluated with concrete benchmark scenarios over synthetic networks. We also use the framework to assess the level of common interests among sample users of a popular online social application. Our results confirm that clusters formed by user selection have low degrees of similarity; our framework could, hence, be valuable in forming communities with higher coherence of interests.**

## I. Introduction

The term *community* denotes a social group of people that have one or more things in common. Whether this is residence, geographical neighborhood, traditions, or interests and ideals, communities have been long attracting the interest of sociologists and psychologists thanks to their potential to motivate and shape human behavior. On the contrary, *virtual communities* have emerged more recently and, almost always, transcend distance barriers. Empowered by the Internet, these online communities socialize in virtual spaces provided by social networking sites. A major research question is then how could the dynamics of these virtual worlds be exploited for more efficient design of networked communication protocols and which factors may shape the end-user (the network communication subject) behavior. It has been reported, for example, that higher similarity in the interests/preferences of online social group members favors collaborative, and even altruistic, behavior in content replication [10] and content dissemination [1] scenarios. But is such similarity present in social networks, where users tend to select their friends/followers with very different criteria, including acquaintance, social status, educational and family background? To answer this question, we need to devise mechanisms and tools that can assess the similarity of interests among social group members and leverage the structure this similarity embeds in their social network.

Our work in this paper addresses this requirement by poring over the interest-based community detection problem. We propose a framework, which we call "*ISCoDe*", for assessing the similarity in online social communities (Fig. 1). Input to *ISCoDe* are the interests of the communities' member nodes in certain thematic areas, hereafter called "interest classes", such as music, sports, art. Each interest class could further be split into subcategories (*tags*). In section IV, we give an example of how the end-user interests can be inferred out of a real social network application. *ISCoDe* then proceeds in two steps. First, it quantifies the interest similarity between node pairs through the use of interest similarity metrics. Outcome of this step is a weighted graph representation of the social network, with edge weights corresponding to the similarity metric values. In a second step, *ISCoDe* can invoke standard community detection algorithms for weighted graphs (for example, [14] [4]) to group nodes into disjoint clusters, connected internally by high-weight edges and to other subsets' nodes with small- or zero-weight edges. These algorithms also assess, in the same time, the quality of this grouping through the modularity metric [16].

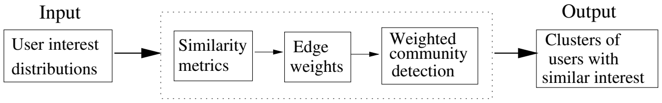

Input
User interest distributions
Similarity metrics
Edge weights
Weighted community detection
Output
Clusters of users with similar interest

Figure 1. The *ISCoDe* framework

We call *ISCoDe* a framework since there are more than one options for its two main processing steps (namely, the derivation of the weighted graph edges and the community detection algorithm). Part of our work, hence, is devoted to the analysis and assessment of these options. For the derivation of graph edge weights, we consider two metrics: the Proportional Similarity [20] and the inverse of the symmetrized Kullback-Leibler divergence [12]. Effectively, each metric could be seen as a different *transformation* from one data set (distribution of user interests over interest classes) to another (graph edge weights). Comparing the outcomes of *ISCoDe* under synthetic user interest distributions, we show that the choice of the similarity metric affects both the sensitivity and the resolution properties of our framework. Note that the similarity metrics we consider are different from the similarity indices that only capture structural equivalence, *i.e.*, same profile of relations to all other nodes in the network, such as the Pearson correlation and the Jaccard coefficient [5].

*Related work*: In the literature, algorithms for detecting community structure have largely been applied to a given network structure, usually modeled as a graph. The most prominent algorithm thereof is that of Girvan and Newman [16], which is highly efficient and overcomes many shortcomings of previously proposed algorithms, such as graph partitioning (*e.g.*, spectral bisection [17], Kernighan-Lin algorithm [11]) and hierarchical methods (*e.g.*, Euclidean distance single linkage clustering) [8]. These methods are not ideal for analyzing general network data since usually it is not known in advance in how many communities the network should be split into and which is the best division. Newman further proposed a simple mapping from a weighted network to an unweighted multigraph and proposed an algorithm for detecting communities in weighted networks [14]. The graph edge weights introduce another set of variables in the community detection process and it is shown in [7] that they can have big influence on the resulting community structure, especially on dense networks.

*Contribution of this paper*: As in Newman's approach, current practice in community detection consists in applying modularity-maximizing clustering algorithms over *given* (weighted) graphs. Our work has a different starting point. In our paper the network structure, *e.g.*, edge weight set, is not given beforehand. It is rather generated by *ISCoDe* out of the distributions of user interests over different thematic areas, the only information we assume known and given to us. Since our framework uses the interest distributions as its input for community detection, in the same time it becomes a means of assessing the effectiveness of similarity metrics that carry out the *mapping* of interest distribution differences. This paper, hence, explores how effectively different *mappings* facilitate the detection of the underlying interest similarity structure when the "commodity" community detection algorithms are applied on their *images*, *i.e.*, the weighted edge sets they generate. Through the application of the framework and the presented results, the effectiveness of the proposed framework that advocates projecting distributional differences to a weighted graph and using commodity approaches for identifying communities thereof, is assessed and established.

The paper proceeds as follows: Section II describes briefly the scope and processing steps of the *ISCoDe* framework for interest-based clustering. The evaluation methodology and experimental results are presented in Section III. An application to a real network is described in Section IV. Finally, we summarize the major conclusions of the paper in Section V and point to interesting problems for future work.

## II. THE *ISCoDe* FRAMEWORK FOR DETECTING COMMUNITIES OF NODES WITH SIMILAR INTERESTS

In general, we want *ISCoDe* to satisfy three main requirements:

**Correctness:** The framework should be able to distinguish correctly existing community structure. Whereas it may not always be possible to conclude whether such structure really exists, the outcome of the framework should at least agree with our intuition in certain benchmark scenarios, where this structure is evident.

**Sensitivity:** The framework should be able to adapt to changes of the user interest distributions and reflect the strength of the community structure.

**Resolution:** The framework should be able to identify important community structure irrespective of its scale and the overall network size.

We evaluate *ISCoDe* along these lines in section III. In the rest of this section, we detail the two processing steps of the framework and present baseline choices for populating them.

### *A. From interest distributions to the weighted graph*

Let $\mathcal{N} = \{1, 2, \ldots, N\}$ be the set of the network nodes (online social network users) and $\mathcal{M} = \{1, 2, \ldots, M\}$ the set of interest areas (classes). We assume that for each node $n$ we can have an estimate of $F^n$, the probability distribution of its preferences over the $M$ interest areas, which takes discrete values $F_1^n, F_2^n, \ldots, F_M^n$ with $\sum\limits_{m \in \mathcal{M}} F_m^n = 1$. Practically, $F_m^n$ could be measured through the normalized request rate of node $n$ for data objects (content) of type $m$ or some other form of interest expression in a certain area (*e.g.*, subscription to this category's tags). In section IV, we describe this process for a particular online social application.

From the node interest distributions, we can then compute the pairwise similarity in the interests of two nodes drawing on measures of distributional similarity. Hereafter, we describe and focus on two of the possible choices: a) the Proportional Similarity (PS) metric, which is shown in [20] to satisfy 11 criteria suggested as suitable for a measure of similarity between distributions; and b) the inverse of Kullback-Leibler symmetrized divergence (InvKL) [12]. InvKL projects the difference between two interest distributions to a significantly broader range of values compared to the PS metric, *i.e.*, $(0, +\infty)$ *vs.* $[0, 1]$, thus shaping the resolution properties of the framework, as we will see later in Section III.

*1) Proportional Similarity (PS) metric:* With the PS metric, the interest similarity $PS_{F^i,F^j}$ between two nodes $i$ and $j$, with interest distributions $F^i$ and $F^j$, equals [20]:

$$PS_{F^i,F^j} = 1 - \frac{1}{2} \sum_{m=1}^{M} \left| F_m^i - F_m^j \right| . \tag{1}$$

*2) Inverse KL (InvKL) symmetrized divergence:* Our second metric is the inverse of the Kullback-Leibler (KL) symmetrized divergence, a metric capturing the distance between two distributions

$$InvKL_{F^i,F^j} = \big( \sum_{m=1}^{M} F_m^i log \frac{F_m^i}{F_m^j} + F_m^j log \frac{F_m^j}{F_m^i} \big)^{-1}. \tag{2}$$

The InvKL metric takes values in $(0, +\infty)$. The KL divergence goes to infinity in cases where there is no interest in one interest category from one node, whereas there is non-zero interest in it from another. In order to avoid such problems, smoothing methods (*e.g.*, interpolation and backing-off schemes) can be used. These have been studied in statistical language modeling in order to estimate the distribution of natural language as accurately as possible [13]. In our case non-zero request rates for interest classes can be discounted with different discounting methods (see [13]), whereas interest classes for which there is no interest can be given a small $\epsilon > 0$ probability.

### *B. From weighted graphs to communities*

Out of the full population of clustering algorithms, relevant to our objectives are those carrying out density-based graph clustering [3]. Namely, they take as input a graph and partition it in a way that some notion of density (in our case: the weights of intra-cluster edges) is significantly higher within a partition than across different partitions (inter-cluster edges). Within the complex networking community the de-facto criterion for assessing the quality of the partitioning is modularity [14], [16]. Modularity sums across all partition clusters the fraction of within-cluster edges minus the expected fraction of edges that would fall within the same cluster were they selected at random. For a given partition of a weighted graph $G(V, E)$, where $V$ is the set of network nodes and $E$ the edge set capturing pairwise interest similarities, modularity $Q$ equals [14]

$$Q = \sum_{c=1}^{C} \left[ \frac{l_c}{L} - \left( \frac{d_c}{2L} \right)^2 \right], \tag{3}$$

where the sum is over the $C$ communities of the partition, $L$ is the sum of the weights of all edges in the graph, $l_c$ is the sum of weights over edges lying fully within community $c$, and $d_c$ the respective sum over the full set of edges incident to nodes in $c$. Modularity takes values in the interval $[-1/2, 1]$ [2]. It becomes zero for community structures that do not differ than what one would get by random chance, whereas values above $0.3 - 0.4$ suggest strong community structure.

Our framework lends to the use of different modularity-maximization algorithms. One example is the divisive clustering algorithm Newman proposed in [14] for weighted graphs. The algorithm iteratively removes from the original graph the edge with the highest "edge betweenness" (defined as the number of shortest paths between pairs of nodes traversing the edge) and recalculates modularity and edge betweenness values till modularity does not increase any further. The complexity of the algorithm is $O(|E|^2|V|)$, which for dense graphs yields $O(|V|^5)$.

More generally, the problem of finding a partition that maximizes modularity in general graphs has been formulated as an Integer Linear Program (ILP) and shown to be NP-hard [2]. Proposed heuristic algorithms for modularity maximization draw on simulated annealing [19] or extremal optimization [6]. More commonly used and computationally friendlier, however, is the greedy agglomerative clustering algorithm of Clauset *et al.* [4], [15]. We simply extend it to weighted graphs by directly relating it with the definition of modularity in weighted graphs in (3). Initially each vertex is viewed as a discrete cluster of size one. The algorithm then iteratively merges the two clusters that yield the largest modularity increase. The algorithm completes in at most $|V| - 1$ steps and has an implementation cost of $O(|V|^2 log|V|)$ [2] permitting scalability for large graph sizes. We retain the greedy algorithm as the baseline for the assessment of *ISCoDe* in Section III-A.

## III. *ISCoDe* EVALUATION

We work with synthetic networks of $N$ member nodes with *controllably* similar interests in order to evaluate the correctness, sensitivity, and resolution properties of the framework. With modularity as the fitness metric of the detected community structure, structures featuring tighter communities with cleaner separation from each other should see higher $Q$ values than equinumerous yet "looser" structures. Moreover, with respect to *ISCoDe*'s resolution, we recall the remarks by Fortunato and Barthèlemy in [9] that algorithms seeking to maximize modularity may fail to identify important structures smaller than a scale. In concluding whether the identification of further distinct communities within a single one is meaningful, we adopt the weak "community" condition by Radicchi [18], *i.e.*, a community $c$ is correctly identified as one if

$$\frac{l_c}{L} - \left( \frac{d_c}{2L} \right)^2 > 0. \tag{4}$$

Note that in *ISCoDe* the resulting modularity values are significantly affected by the choice of the similarity metric. Contrary to other studies in literature, where community detection algorithms maximizing modularity are studied on given complex weighted graphs, *ISCoDe* adds the additional transformation step of interests to graph edge weights. Therefore, another requirement from the evaluation process is to show how the two interest similarity metrics affect the three framework requirements.

In the general setting, the network population is organized into $k$ groups. Each group is interested in $M$, generally different, interest classes, which are the same for all member nodes of a given group. We form $k$ equal-size groups of $N/k$ users: nodes $1..N/k$ are assigned to group 1, nodes $N/k + 1..2N/k$ to group 2, and so on (for the sake of the example, we take $N/k$ to be an integer). We then control the similarity within and across the $k$ groups as follows:

**Interest similarity *across* groups.** This is controlled in two ways. Firstly, through the number of common interest areas between groups, which may take any value $r$ in $[0, M]$. Secondly, and this relates to the way the similarity *within* a single group is controlled, through the way the interests overlap. We consider two scenarios for the distribution of common interests between two groups: a) the $r$ common interest areas are simultaneously the $r$ least interesting for group $g$ and the $r$ most interesting for group $g+1$, $0 < g < k$ (*L(ast)-F(irst)*, Table I(a)); b) the $r$ common interest areas are the $r$ most interesting for the users of all $k$ groups (*F(irst)*-

*F(irst)*, Table I(b)). These scenarios present two extreme cases regarding the interest similarity across groups. Given that the number of common interest areas and the distributions are held fixed, the L-F(F-F) scenario yields the smallest(highest) similarity.

Table I
EXAMPLE OF THE TWO INTEREST-OVERLAP SCENARIOS FOR $k = 3$, $M = 5$. THE ORDER OF INTEREST CLASSES MARKS ALSO THE ORDER OF NODES' INTERESTS WITHIN A GROUP.

(a) L-F with a single overlap interest class ($r = 1$)

| Group 1 | Group 2 | Group 3 |
|---|---|---|
| 1 | 5 | 9 |
| 2 | 6 | 10 |
| 3 | 7 | 11 |
| 4 | 8 | 12 |
| 5 | 9 | 13 |

(b) F-F with two overlap interest classes ($r = 2$)

| Group 1 | Group 2 | Group 3 |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 2 | 2 |
| 3 | 6 | 9 |
| 4 | 7 | 10 |
| 5 | 8 | 11 |

**Interest similarity *within* groups.** The interests of nodes within a group are spread over the ordered $M$ interest classes inline with the Zipf distribution[1]. The skewness parameter $s$ of the distribution differs for each group node. The interest of the first node of each group are uniformly distributed ($s_1 = 0$) and $s$ increases with constant step $p$ so that for node $n$, $s_n = p(n-1), p \in \mathcal{R}$. Higher $p$ values increase the skewness in the interest distribution and concentrate the node interests' mass in the higher-order interest classes. Interestingly, changes of $p$ also affect the similarity of interests between nodes belonging to different groups depending on the overlap scenario (Table I): higher $p$ values result in weaker (stronger) inter-group similarity in the $L-F$ ($F-F$) overlap scenario.

In summary, by calibrating $p$, the overlap scenario and the number of common interest classes, we can produce networks with community structures of variable discernibility.

### A. Experimental results and discussion

We show and discuss representative results from our experimentation with *ISCoDe* on synthetic networks that outline the main behavior of the framework. All experiments are carried out with the greedy agglomeration algorithm in [4] since it yields significantly faster run times than its competitors[2].

*1) Correctness and sensitivity experiments:* In this set of experiments, $N = 80$ and $k = 4$. The impact of $M$ was found to be minimal, thus we present herein results only for $M = 20$. We vary the interest overlap scenarios (as in Table I), the number of common interest classes, one (Tables II(a), II(c)) or half of them (Tables II(b), II(d)), and the skewness of the interest distributions, smaller $p$ values representing more uniform distributions within a single group nodes.

The first remark is that both metrics produce the same intuitive community partitions in the presence of strong community structure, as in Tables II(a) and II(c) for low $p$ values. On the contrary, when such structure is not evident, the two metrics result in considerably different partitions.

The second remark has to do with the higher sensitivity of the framework when the PS metric is used in its first processing step. The modularity of the resulting partitions under the PS metric decreases when the interests of nodes are more randomly diffused over the different interest classes and goes down to zero when the similarity structure tends to disappear, as in Table II(d). On the contrary, the resulting modularity values under the InvKL metric are almost insensitive to the changes in the input interest distributions. With InvKL the modularity values are dominated by the edge weights between individual node pairs; these tend to be very high ($\gg 1$) for highly similar nodes and very low for highly dissimilar nodes. Finally, as $p$ increases, the interest distributions of most nodes tend to be more concentrated on the first group objects, and the interest distributions become less uniform. For cases shown in Tables II(a) and II(b), this results in increasing modularity under the PS metric, thanks to the decreasing weights of inter-group edges, *i.e.*, nodes initially assigned to different groups. It has the opposite effect for cases shown in Tables II(c) and II(d), where increasing $p$ leads to stronger ties between nodes in different groups. On the contrary, InvKL does not adapt to any of these changes.

*2) Resolution experiments:* We run two additional experiments focusing on the impact of the two similarity metrics upon the overall framework resolution. The first experiment involves nodes with *highly similar interests*. All nodes are interested in the same $M$ objects, in the same order. They differentiate only slightly in how their interests are spread over the $M$ interest classes, modelled by Zipf(s) distributions with $s$ varying from 0 to its maximum value in steps of $p = 0.01$. The second experiment involves nodes with *highly dissimilar interests*; there is a single common interest class between successively ordered nodes. The experiment resembles the L-F overlap scenario shown in Table I(a), if each group contained only one node. The results from the two experiments are reported in Table III and clearly demonstrate the capacity of the two metrics to illuminate better different parts of the interest similarity range.

Mapping highly similar interest distributions to a much broader edge weight value range (Figure 2(a)), InvKL can resolve more communities than PS in the first experiment, all of which satisfy Radicchi's weak community condition of (4). On the contrary, PS tends to group small communities together. Notably, the communities produced by both metrics do not satisfy the inequality

$$l_c < \sqrt{2L}, \qquad (5)$$

which according to [9] suggests that community $c$ may be the combination of two or more smaller communities that cannot simply be detected when pursuing the optimization of modularity due to their small size.

The situation is reversed in the second experiment: it is now PS that can recognize smaller communities, as shown in Table III(b). Moreover, (5) is satisfied, implying that there

[1] Zipf distributions have been used in the recent past to capture preferences for web objects. Furthermore, they exhibit high modelling simplicity and flexibility, in that proper manipulation of their single parameter $s$, gives rise to a wide set of distributions ranging from uniform ($s = 0$) to highly skewed ones with power-law characteristics ($s >> 0$).

[2] We run experiments also with the divisive clustering algorithm in [14] but we had to restrict to small group sizes. In these cases, we obtained similar results with respect to community structure and modularity values.

Table II

CORRECTNESS AND SENSITIVITY EXPERIMENTS: MODULARITY AND COMMUNITIES FORMED FOR DIFFERENT VALUES OF $p$, $N = 80$, $M = 20$

(a) L-F, 1 common object

| | PS | | | InvKL | | |
|---|---|---|---|---|---|---|
| | $Q$ | $C$ | partition | $Q$ | $C$ | partition |
| $p = 0.02$ | 0.6849 | 4 | {1..20}...{61..80} | 0.7498 | 4 | {1..20}...{61..80} |
| $p = 0.04$ | 0.6925 | 4 | {1..20}...{61..80} | 0.7493 | 4 | {1..20}...{61..80} |
| $p = 0.06$ | 0.6992 | 4 | {1..20}...{61..80} | 0.7483 | 4 | {1..20}...{61..80} |
| $p = 0.08$ | 0.7048 | 4 | {1..20}...{61..80} | 0.7698 | 8 | {1..10}...{71..80} |
| $p = 0.10$ | 0.7095 | 4 | {1..20}...{61..80} | 0.7745 | 8 | {1..10}...{71..80} |

(b) L-F, $M/2$ common objects

| | PS | | | InvKL | | |
|---|---|---|---|---|---|---|
| | $Q$ | $C$ | partition | $Q$ | $C$ | partition |
| $p = 0.02$ | 0.3594 | 2 | {1..40} {41..80} | 0.7667 | 4 | {1..20}...{61..80} |
| $p = 0.04$ | 0.3669 | 2 | {1..40} {41..80} | 0.7490 | 4 | {1..20}...{61..80} |
| $p = 0.06$ | 0.3756 | 2 | {1..40} {41..80} | 0.7475 | 4 | {1..20}...{61..80} |
| $p = 0.08$ | 0.3938 | 3 | {1..20} {21..40} {41..80} | 0.7687 | 8 | {1..10}...{71..80} |
| $p = 0.10$ | 0.4142 | 3 | {1..20} {21..40} {41..80} | 0.7730 | 8 | {1..10}...{71..80} |

(c) F-F, 1 common object

| | PS | | | InvKL | | |
|---|---|---|---|---|---|---|
| | $Q$ | $C$ | partition | $Q$ | $C$ | partition |
| $p = 0.02$ | 0.5711 | 4 | {1..20}...{61..80} | 0.7498 | 4 | {1..20}...{61..80} |
| $p = 0.04$ | 0.5146 | 4 | {1..20}...{61..80} | 0.7492 | 4 | {1..20}...{61..80} |
| $p = 0.06$ | 0.4465 | 4 | {1..20}...{61..80} | 0.7480 | 4 | {1..20}...{61..80} |
| $p = 0.08$ | 0.3734 | 4 | {1..20}...{61..80} | 0.7692 | 8 | {1..10}...{71..80} |
| $p = 0.10$ | 0.3038 | 4 | {1..20}...{61..80} | 0.7730 | 8 | {1..10}...{71..80} |

(d) F-F, $M/2$ common objects

| | PS | | | InvKL | | |
|---|---|---|---|---|---|---|
| | $Q$ | $C$ | partition | $Q$ | $C$ | partition |
| $p = 0.02$ | 0.1103 | 4 | {1..20}...{61..80} | 0.7496 | 4 | {1..20}...{61..80} |
| $p = 0.04$ | 0.0841 | 4 | {1..20}...{61..80} | 0.7481 | 4 | {1..20}...{61..80} |
| $p = 0.06$ | 0.0610 | 4 | {1..20}...{61..80} | 0.7444 | 4 | {1..20}...{61..80} |
| $p = 0.08$ | 0.0422 | 4 | {1..20}...{61..80} | 0.7611 | 8 | {1..10}...{71..80} |
| $p = 0.10$ | 0.0485 | 5 | {1..15}...{61..75} {16..20,36..40,56..60,76..80} | 0.7549 | 8 | {1..10}...{71..80} |

Table III

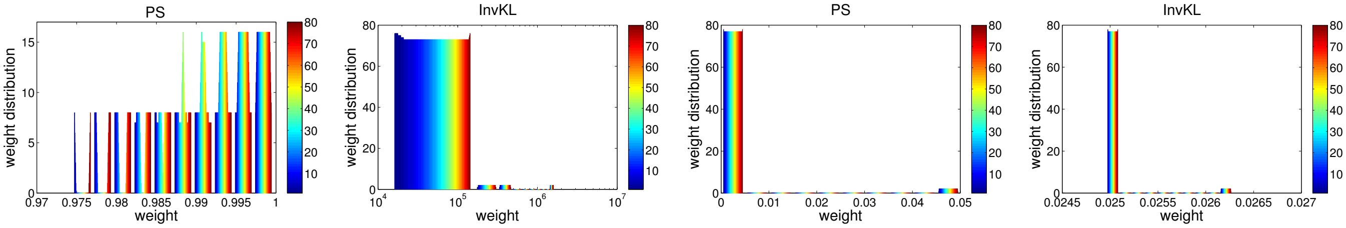
RESOLUTION EXPERIMENTS: MODULARITY AND COMMUNITIES FORMED, $N = 80$, $M = 20$

(a) Similar nodes

| PS | | | InvKL | | |
|---|---|---|---|---|---|
| $Q$ | $C$ | partition | $Q$ | $C$ | partition |
| 0.0215 | 2 | {1..38} {39..80} | 0.6740 | 5 | {1..14} {15..28} {29..44} {45..61} {62..80} |

(b) Dissimilar nodes

| PS | | | InvKL | | |
|---|---|---|---|---|---|
| $Q$ | $C$ | partition | $Q$ | $C$ | partition |
| 0.7860 | 10 | {1..8}..{73..80} | 0 | 1 | {1..80} |



(a) Similar nodes

(b) Dissimilar nodes

Figure 2. Resolution experiments: Edge weight distributions

are more non-detected communities. InvKL, on the contrary, *cannot* since it squeezes all edge weight values that result from the first processing step within an interval of 0.012 width (Figure 2(b)).

However, an important question is regarding the level of resolution, *i.e.*, in which cases communities should be more resolved. Intuitively, it seems more important to identify finer community structure in a network with more similar nodes, than in case of dissimilar ones. Hence, the resolution advantage of InvKL at high similarity scenarios may overweigh its disadvantage at low similarity ones.

## IV. APPLICATION TO A REAL NETWORK

We apply *ISCoDe* to data traces extracted from the Delicious website (www.delicious.com). Delicious is a social bookmarking application where users can save all their web bookmarks (annotated with tags) online, share them with other users, and track what other users are bookmarking themselves. Each user forms a network with other users that have subscribed to see their bookmarks. We use the organization of users into networks and interests into tags to generate the user interest distributions and feed *ISCoDe* with them.

**From user interest profiles to interest distributions.** Let $M$ be the set of most popular tags used by each Delicious user. Let $B_m^n$ be the number of bookmarks tagged with $m$ ($1 \leq m \leq M$) by user $n$ ($1 \leq n \leq N$). Then the (normalized) interest of node $n$ in tag $m$ is given by the ratio of the number of bookmarks tagged with $m$ by node $n$ over the total number of bookmarks of this user:

$$F_m^n = \frac{B_m^n}{\sum_{m=1}^{M} B_m^n}. \qquad (6)$$

**Experimentation set-up.** The Delicious network is crawled in two ways. The first method starts from four Delicious accounts (root users) chosen randomly on the website. From each root user 29 users – who follow the root user – are extracted using a breadth-first exploration of the graph formed by these links. To avoid the long tail of infrequently used tags, only bookmarks that contain the 99 most popular tags are considered for each user. The interest profiles of 120 in total users are derived from (6). Running the community detection algorithm with the PS and InvKL metrics results in community structures with modularity values $0.0672$ and $0.1130$, respectively. The second procedure is similar to the first one; only now the four root users are selected *among those having placed recent bookmarks on the website*. Here the 30 highest preference tags (as derived from (6) are kept for each user. The PS and InvKL metrics for this case result in modularity values $0.0754$ and $0.1971$, respectively.

Our results suggest that modularity values are higher in the second case, as expected, since many users are interested in the same tags. Overall, however, the modularity values of the respective users' partitions are low, implying that Delicious user networks do not display interest similarity structure. Users do not follow other users based on similarity of their tagged bookmarks. We argue that the satisfaction of users from the Delicious network would greatly increase if users formed communities taking higher account of their interests' matching. The framework proposed in this paper could be valuable in this respect.

## V. Conclusion and future work

In this paper we proposed a framework, *ISCoDe*, for clustering of users (nodes) according to their interests. The framework is straightforward and can be used as a guide for the formation of more interest-coherent communities in online social networks. It consists of two steps. First, it quantifies the similarity in the interests of network members with the help of interest similarity metrics. These metrics become the edges of the weighted graph that models the social network. Then a community detection algorithm is applied to this graph to extract communities of nodes with similar interests.

We have investigated two similarity metrics, PS and InvKL, for the derivation of weighted graph edges from user preferences. Our results suggest that both metrics produce reasonable partitions for strong community structure. However, the InvKL metric is not as sensitive as PS regarding the changes in the strength of community structure. On the other hand, InvKL has a higher resolution in networks of nodes with highly similar interests. In any case, our paper has illustrated that the interest distribution mappings influence the discernibility of the framework, by emphasizing or de-emphasizing differences in the interest distributions. This insight can be useful in the search for other mappings that map more effectively distribution differences into values better "matched" to the commodity community detection machinery in the sense that they result in more effective community detection.

## VI. Acknowledgments

We thank Walter Colombo and Martin Chorley from Cardiff University for kindly providing us with traces they collected from the Delicious network. This work has been supported in part by the European Commission IST-FET project SOCIALNETS (FP7-IST-217141) and the Marie Curie grant RETUNE (FP7-PEOPLE-2009-IEF-255409).

## References

[1] S. M. Allen, G. Colombo, and R. M. Whitaker. Cooperation through self-similar social networks. *ACM Trans. Auton. Adapt. Syst.*, 5(1):4:1–4:29, February 2010.

[2] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *IEEE Trans. on Knowl. and Data Eng.*, 20:172–188, February 2008.

[3] U. Brandes and T. Erlebach. *Network Analysis: Methodological Foundations (LNCS)*, volume 3418. Springer, Heidelberg, March 2005.

[4] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111+, December 2004.

[5] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Routledge Academic, 2 edition, January 1988.

[6] J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72(2):027104+, August 2005.

[7] Y. Fan, M. Li, P. Zhang, J. Wu, and Z. Di. The effect of weight on community structure of networks. *Physica A: Statistical Mechanics and its Applications*, 378(2):583–590, May 2007.

[8] A. Fernández and S. Gómez. Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. *J. Classif.*, 25:43–65, June 2008.

[9] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proc Natl Acad Sci (USA)*, 104(1):36–41, January 2007.

[10] E. Jaho, M. Karaliopoulos, and I. Stavrakakis. Social similarity as a driver for selfish, cooperative and altruistic behavior. In *4th IEEE WoWMoM Workshop on AOC'10*, Montreal, Canada, June 2010.

[11] B. W. Kernighan and S. Lin. An Efficient Heuristic Procedure for Partitioning Graphs. *The Bell system technical J.*, 49(1):291–307, 1970.

[12] S. Kullback. *Information theory and statistics*. J. Wiley and Sons., 1959.

[13] R. D. Mori. *Spoken Dialogues with Computers*. Academic Press, Inc., Orlando, FL, USA, 1997.

[14] M. E. J. Newman. Analysis of weighted networks. *Phys. Rev. E*, 70(5):056131, November 2004.

[15] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69:066133, September 2004.

[16] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113+, Feb. 2004.

[17] A. Pothen, H. D. Simon, and K. P. Liou. Partitioning Sparse Matrices with Eigenvectors of Graphs. *SIAM J. Matrix Anal. Appl*, 11(3):430–452, May 1990.

[18] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proc Natl Acad Sci USA*, 101(9):2658–2663, March 2004.

[19] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1):016110+, July 2006.

[20] J. Vegelius, S. Janson, and F. Johansson. Measures of similarity between distributions. *Quality and Quantity*, 20(4):437–441, December 1986.