Towards Maximising Cross-Community Information Diffusion

Václav Belák, Samantha Lam, and Conor Hayes
DERI, NUI Galway, IDA Business Park, Lower Dangan, Galway, Ireland
Email: vaclav.belak@deri.org, samantha.lam@deri.org, conor.hayes@deri.org

***Abstract*—In recent years, companies have started to utilise online social communities as a means of communicating with and targeting their employees and customers, and such online communities include discussion fora. The conversational dynamics of users in fora can influence their neighbours in the underlying social network. We make use of such influence to target specific communities with information, i.e. post in them, because a post is generally shared with the community and not just with individual users. In short, we study information diffusion across communities and show that we can achieve high community (and user) spread using an efficient targeting strategy. In order to achieve this, we use a set of novel measures for cross-community influence and show that it outperforms other targeting strategies on two different data-sets: the largest Irish online discussion system, Boards.ie, and technical support fora, SAP SCN.**

I. Introduction

Online communities have become increasingly important in the context of many services provided on the Internet. In particular, many companies have started to utilise online social communities as a means of communicating and targeting their customers and other partners. However, in order to exploit the full potential the communities offer to their stakeholders, an efficient communication strategy has to be employed. It is not surprising to observe that if users are continuously flooded by a torrent of new stimuli, they may become increasingly inert to any further provocation. Thus, it is of utmost importance to carefully select a strategy of who and how to approach such that the expected outcome of network coverage is maximised. This maximisation problem is further compounded by the fact that social networks and communities are inherently dynamic. As a result, the study of spreading behaviour across networks has garnered much inter-disciplinary interest in recent years, from the spread of disease in a population [1], [2] to the spread of influence through a social network [3].

In the case of discussion fora, the scenario is somewhat different to information spreading from a set of seed actors because they represent a different setting; a message is shared with *all* participants in the forum (i.e. the forum's community). Thus, the problem becomes how to target a message to engage a *set of users* rather than specific, individual users in a network, such that the message reaches as many users in the network as possible, i.e. *user spread* is maximal. Furthermore, since fora are centred around specific interests, the problem can be formulated alternatively as a maximisation of the spread of a message across as many *communities* as possible, i.e. *community spread* is maximal. High spread over communities corresponds to a high diversity or heterogeneity of the resulting set of actors that have adopted the message. The **main problem** we address is therefore formulated as a *prediction* of the set of communities to target such that the user and community spreads are maximised in the future.

In our recent work, we first provided a framework to identify influential communities [4], and then gained some first insights in the problem of maximisation of user spread [5]. In this paper we extend [5] to maximise *community spread* as well as user spread and evaluate it on two characteristically different data-sets by conducting four experiments with two diffusion models. We compare the communities chosen under the framework with those identified as 'central' groups as defined by [6]. We then target combinations of the most influential/central groups according to these metrics and evaluate their *user* and *community* spreads.

Our **main aim** is to investigate robustness of our framework with respect to: a) the character of the system; our first data-set represents a general-purpose discussion fora, while the other is a business system for technical support, b) the dynamics of influence simulated by two distinct diffusion models, and c) its ability to achieve high spreads over both users and communities. Overall, our framework achieved higher user and community spreads efficiently (i.e. for small numbers of targeted communities) in both data-sets in both diffusion models.

In short, the **main contributions** of this paper are that we:

- Extensively investigate the suitability and robustness of our framework to predict influential communities based on the underlying social (communication) networks of Boards.ie and SAP SCN.[1] In particular we show the efficacy of the framework in both user and community spreads.
- Compare the differences between the two systems; we found that while Boards.ie is very reciprocal system in which information is likely to propagate efficiently, targeting communities in SAP SCN is much harder due to particular properties of the underlying social network. As a result, our findings may help the stakeholders to manage the social capital of their system more efficiently by enhancing their communication strategy.

[1] See http://boards.ie and http://scn.sap.com.

Please note that we use the term *impact* as a means of quantifying the phenomenon of *influence*. We refer to *information flow* as the observed process of influence. In particular, we note that although the notion of influence in the context of social media analytics refer to the ability of an actor to change behaviour of its neighbours [7], in this paper our definition of community influence is specifically tied to the reply-to activity.

The remainder of the paper is organised as follows. In the next section we refer to the related work. In Section III we summarise the major concepts of the framework itself together with the data-sets, their preparation, and the diffusion models we used for the evaluation. The experimental setup is clarified in Section IV and the results of the experiments are then presented in Section V. The last section concludes the paper and outlines our aim for future work. Finally, the data-sets and software we used to conduct our experiments, along with an additional supplementary material, are available online.[2]

II. Related Work

In this section we first refer to the related research of information flow and conversational dynamics in discussion fora. In the second part, the models of information diffusion are discussed.

A. Information Flow in Discussion Fora

McGlohon and Hurst [8] examined the flow of information in USENET. A specific feature of USENET is that it is possible to send or forward a message to multiple newsgroups — to *cross-post* it. As a cross-posted message belongs to multiple groups, they developed a thread-ownership model based on the notion of author-group *devotedness* of the users measured by the distribution of their activity. In our proposed framework, we draw upon their approach and measure the devotedness in a similar manner. However, although there is no explicit cross-posting in online fora, its users can and do post in multiple fora which then receive replies from members of other fora.

Wu et al. [9] modelled the flow of information in discussion fora using its reply-to network as a proxy. The authors used a PageRank-inspired random walk model to show how multiple topics flow across discussion threads, and to predict future interests of the users based on their conversational activity. They define a user as participating in a discussion if the user posts at least once in it, and information 'flows' from the user being replied to. We adopt this notion of information flow, which is also similar to how Song et al. [10] define it for personalised recommendation. However, their approach assumes that information 'dilutes' as it flows in that it is not duplicated through propagation. Reply-to relations were also found to have many similar properties to classic friendship relations and could be used in the prediction of user grouping behaviour [11].

The problem of finding influential actors within a social network has been intensively studied in social network analysis [12], but less so on the level of communities. For the individual actors, a classic approach is to use a centrality measure like actors' degrees. Everett and Borgatti [6] generalised several centrality measures to groups of actors. For instance, they defined *group degree centrality* as "the number of non-group nodes that are connected to group members". Hence the group degree captures the relation between a group of actors and the *rest* of the network but not between two or more groups. Their measure thus extends the traditional actor degree heuristic to a community level by aggregating the users' degree into one actor.

In our previous paper [4] we presented a framework for cross-community influence analysis. In this paper, we use that framework to develop a technique for identifying influential communities to efficiently stimulate communication in fora.

B. Diffusion Models

Several models of how information or an action diffuses over a social network has been proposed (see e.g. [7] for a recent survey). The problem of maximising the spread of information or influence was first introduced by Kempe et al. [3], who considered two generalisations of many previously defined models — *Independent Cascade Model* (ICM) and *Linear Threshold Model* (LTM).

Both models consider a social network represented by a directed weighted graph $G = (V, E)$, where nodes V are the actors and a weighted edge $w_{i,j} \in E \subseteq V \times V$ expresses a propensity of an actor j to adopt information or an action from i. Each actor can be either *active* or *inactive* and both models proceed in discrete steps where the activation spreads from active to non-active ones. The main difference between the two is the concrete mechanism of the spreading process.

In the case of LTM, a non-active node j at iteration t is activated if $\sum_{i \in AN_i} w_{ij} \geq \theta_j$, where AN_i is a set of active neighbours of j in the previous iteration $t-1$, and θ_j is a threshold expressing how many neighbours of j have to be active in order to activate it. That is, the decision of whether a node becomes active or not depends *only* on the weighted sum of its active neighbours and its threshold. The previously activated nodes remain active in the following iterations and the diffusion process unfolds *deterministically* until the process converges, or until the maximum number of iterations has been reached.

Without any loss of generality, it is commonly required that $\forall j \in V : \sum_{i \in V \wedge i \neq j} w_{ij} = 1$ and thus the weights can be interpreted as *probabilities*, which is the main intuition behind ICM. In this model the diffusion process unfolds *stochastically*. The diffusion process starts again with a set of seed nodes and at each iteration t, each node i that has been activated in a previous iteration $t-1$ has exactly one try to activate each of its non-active neighbours j, and it succeeds with a probability w_{ij}. If multiple neighbours of j are activated in the previous iteration, they attempt to activate it in a random order. Hence the individual attempts are *independent* of each other. If any of the j's neighbours succeeds, it becomes active in the next iteration $t+1$. However, whether any of the j's neighbours succeed or not, they have *no* further chances to

[2] See http://belak.net/doc/2012/asonam.html. The SAP data-set is an *anonymized* version of crawled data from the *public* SAP SCN portal.

spread their activation in the following iterations. The process again stops when it converges or until the maximum number of iterations has been reached.

LTM intuitively corresponds to when the influence of the *neighbourhood* of a node to itself is *aggregated* (since the active nodes influence their neighbours until the end of the simulation), while ICM mimics the case of when there is one influential neighbour at a given time. We used these two models because they were previously successfully used as baselines [3], and because our aim was to investigate robustness of our framework with respect to different biases related to the influence dynamics simulated by the two models.

III. Preliminaries

This section presents the framework we developed in [4] and the data we used. First we describe the data-sets and systems: Boards.ie and SAP SCN ("Boards" and "SAP" hereafter). Next, we describe how the information flow network we use for evaluation was derived from the reply-to network, as well as the diffusion models themselves. Finally, we formally define the notion of cross-community impact and other related measures.

A. Data-Sets

Both Boards and SAP are structured according to themes into *fora*, optionally further into their subfora, and finally into *threads* of *posts* centred around a particular conversation topic. Each post has an author, who can be either a registered *user* or a guest. Since all the guests' posts in Boards are stored with the same user identifier, we omitted them from the analysis. A set of users who have posted at least once to any forum within a certain time-period form a *community* of that forum in the period. Threads have a tree-like structure as one post can be in *reply to* another one. Even though there is no direct way to post a message into multiple fora (i.e. to cross-post it), the users can and do participate in multiple fora and thus information can spread from one forum into another.

The set of users linked by the who-replies-to-whom relation thus forms a directed dynamic graph, as the reply relations change in time. The edges of the graph are weighted by the number of replies from one user to another within a given time period. Table I presents some basic statistics of the analysed data. Please note that there are many more edges per node in Boards than in SAP. That means that the behaviour of Boards users is more conversational than of the users of SAP.

	Boards	SAP
number of snapshots (T)	51	31
number of communities	540	33
mean number of nodes per snapshot	5,298	1,984
mean number of edges per snapshot	26,484	5,609
edges per node (over all snaphots)	36.7	5.6

TABLE I: Elementary statistics of the analysed data-sets.

Time-Window Selection: Our problem is to predict which communities to target in the future based on past/current observations. Rather than aggregating the data up to a specific time slice $t-1$ and then evaluating on the final time slice t, we aim to evaluate our targeting in a more robust manner and consider multiple time snapshots. Thus, we segment the data into T snapshots using a sliding time-window.

As our methods are based on cross-fora posting activity, the window length should capture as much of that activity as possible, yet still fine enough to uncover changes in users' behaviour. Let $\tau(p)$ be a *minimum* time it took an author of post p to contribute a message into another forum, i.e. a *cross-fora posting waiting time*. If the author has not posted to any other fora, then $\tau(p) = \infty$.

In order to find a suitable time-window size, we sampled 10,000 posts and investigated the distribution of $\tau(\cdot)$. In Boards we found that in approximately 84% of the cases a user has posted into another forum within 7 days, while a 14-day period covers 88%. This means that doubling the window size would lead to an increase of only 4% in the coverage of cross-fora posting activity and so we decided to choose a one-week window for our analysis. It was in only 3% of the cases that a user has not posted to another forum at all. Similarly, for SAP we found that a 60-day window covers approximately 49% while 120 days increases the coverage again by only 4%. In contrast with Boards, we observed a much lower level of cross-fora posting activity in SAP — in 40% of the cases a user has not posted to any other forum whatsoever. We believe that the cause may be due to the difference in their utility; while Boards is primarily a place for socialising and discussion of a broad range of topics, the users of SAP may be more focused on a particular topic related to their expertise or problem.

In order to investigate how different targeting strategies affect the diffusion process, we took the last 51 weeks of the Boards data between 19.2.2007 and 10.2.2008. This approximates the last year of our data-set and therefore it is the most recent and reasonably stable representation of the system we have. Similarly, we segmented the SAP data into 31 bi-monthly snapshots between 1.5.2006 and 30.6.2011.

B. Inferring the Information Flow Network

We derive our information flow networks from the reply-to networks in a similar manner to [9] such that if j replies to i then there is an information 'flow' from i to j. This is based on the intuition that if you reply to a message then you would have read its content and therefore gained knowledge from it. Thus, information flows *from* the person that has received a reply. The edges are weighted by the *likelihood* of the flow of information from user i to j, w_{ij}, which is calculated as the number of replies from j to i, r_{ji}, normalised by the total number of replies user j posts:

$$w_{ij} = \frac{r_{ji}}{\sum_{l=1}^{n} r_{jl}}, \tag{1}$$

where n is the total number of users. Figure 1 shows how the flow is reversed when an information flow graph is derived from the reply-to graph. For a detailed example of the network inference and discussion of few alternatives, please see [5].

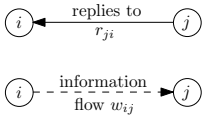

Fig. 1: If j replies to i, it means that information has flowed from i to j. r_{ji} is the number of replies from j to i, w_{ij} is the weight of information flow from i to j.

C. Diffusion Starting From Communities

In the LTM and ICM models described in Section II-B the diffusion process starts from the set of seed nodes and *not* communities. Therefore, it is necessary to extend the models such that the process starts from a set of q targeted *communities*. Since we do not know what the likelihood that a user would respond to a stimulus in a community is, we take a sample of size s in each targeted community to account for as many cases as possible and let the diffusion spread from these users.

The user sampling process itself can be either uniform or it can respect the user's activity in different fora. As we believe respecting the activity is more realistic, we set the probability of a user i to be sampled from community j to $\frac{p_{ij}}{\sum_{l=1}^{k} p_{il}}$, where p_{ij} is the total number of posts of user i in community j, and k is the total number of communities. If the community size was smaller than s, we took all its users.

The extended **Community-Aware Linear Threshold (CALTM)** and **Community-Aware Independent Cascade (CAICM)** models therefore proceed as follows:

1) Select set P of q targeted communities.
2) For each community $C \in P$, sample a set S_C of s users.
3) Obtain a final actor seed set $L = \cup_{C \in T} S_C$. Note that $|L| \leq q \times s$, because the samples of users may overlap.
4) Run the original Independent Cascade or Linear Threshold Model with L as a set of seed nodes.

For the similar reasons as in [3], in the case of CALTM we set the user thresholds θ uniformly at random, which in effect averages over possible values of the thresholds. The main parameters of the extended models are therefore the number of targeted communities q, and the number of users sampled from each targeted communities, s. We investigated up to 5 targeted communities, i.e. $q \in [1, 5]$, and $s \in [1, 50]$ users sampled from each targeted community.

D. Mutual Influence of Fuzzy Communities

We believe any measure of impact between communities should take into account two factors: the degree of *membership* of each user and its *centrality* within each community (the intuition behind this is discussed in [4]). In this section we show how to express and combine these two factors, and how to derive additional measures which are helpful in the interpretation of the cross-community impact.

In order to represent to which communities and to what extent an actor belongs to, let us define an $n \times k$ **membership** matrix $\mathbf{M} : m_{ij} \in [0, 1], \forall\ i : \sum_{j=1}^{k} m_{ij} = 1$ representing the users' affiliations. Columns of $\mathbf{M}$ are fuzzy sets representing the individual communities. $\mathbf{M}$ can be known a priori e.g. from an in-field survey or determined from a community detection algorithm [13]. In our analysis we defined $m_{ij} = \frac{p_{ij}}{\sum_{l=1}^{k} p_{il}}$.

An impact of any given user *within* its communities can be formalised as an $n \times k$ **centrality** matrix $\mathbf{C}$ with elements c_{ij} representing an impact of i-th user to the other users of j-th community. It can be obtained by some centrality measure of a user, e.g. PageRank, in-degree, closeness, etc. We set c_{ij} as the number of replies a user received in a community, which is an *in-degree* of i-th user in a reply-to graph of j-th community. We chose in-degree for our experiments because the reply behaviour is the cornerstone of the conversational dynamics; it is a well-established heuristic for influence maximisation [3] and it has a clear interpretation.

We are now able to formalise the intuition of cross-community impact as a weighted sum of centralities of members of one community within another one:

Definition 1. *We call an* ***impact*** $\mathbf{J}_{ij}$ *of a community i on community j the sum of centralities of the members of i within the community j, weighted by the degrees of membership in i:* $\mathbf{J}_{ij} = \sum_{l=1}^{n} (\mathbf{M}_{li} \times \mathbf{C}_{lj})$.

The $k \times k$ cross-community impact matrix $\mathbf{J}$ can then be obtained as a product of the two matrices: $\mathbf{J} = \mathbf{M}^{\mathrm{T}}\mathbf{C}$. However, social communities usually have different sizes [14], which can bias the impact matrix. A very big community can, from its raw size, accumulate high values in $\mathbf{J}$ despite the fact that its members are not very devoted to it. Therefore we further divide the rows of the impact matrix by the cardinalities of the sets representing the communities — the sum of the columns of the membership matrix — in order to obtain a **normalised impact** matrix:

$$\hat{\mathbf{J}}_{ij} = \frac{\mathbf{J}_{ij}}{\sum_{l=1}^{n} \mathbf{M}_{li}} \tag{2}$$

The normalised impact $\hat{\mathbf{J}}_{ij}$ then represents a weighted *mean* of centralities of members of i-th community in j-th community. The diagonal of $\hat{\mathbf{J}}$ contains self-impact values, i.e. it measures to what extent the highly devoted members of each community are also central in it. If we subtract the diagonal from $\hat{\mathbf{J}}$, we can obtain a vector of communities' **total impact** as row sums:

$$\mathcal{I}(\hat{\mathbf{J}}) = \hat{\mathbf{J}}\mathbf{1} - diag(\hat{\mathbf{J}}) \tag{3}$$

where $\mathbf{1}$ is a column vector of ones of length k.

While some communities may have impact on a relatively small circle of other communities, others may be broadly influential. For instance, a community of system administrators may have an impact to the whole system. Such feature of a community's importance can be characterised as an entropy of the respective row or column of $\hat{\mathbf{J}}$. We further normalise the rows of the matrix in order to obtain probability distributions of impact, i.e. $\hat{\mathbf{J}}^{N}_{ij} = \hat{\mathbf{J}}_{ij} / \sum_{l=1}^{k} \hat{\mathbf{J}}_{il}$. The normalised **importance entropy** of i-th community is then defined as

$$\mathcal{H}_I(i, \hat{\mathbf{J}}) = -\frac{\sum_{m=1}^{k} \hat{\mathbf{J}}^{N}_{im} \log_2 \hat{\mathbf{J}}^{N}_{im}}{\log_2 k} \tag{4}$$

The entropy has range within $[0, 1]$, and because some elements of $\hat{\mathbf{J}}$ may be 0, we define $0 \log_2(0) = 0$ by convention.

The more the impact of i-th community is equally distributed, the more the entropy value is close to 1. We note that in the case of entropy we *include* the diagonal elements (self-impact), because it would differentiates whether the most of the community's impact is concentrated *within* that community or not.

In order to find communities *highly* influencing *many other* communities, we propose to take a product of the total impact (Eq. 3) and its entropy (Eq. 4). While the total impact measures how much one community is capable, *on average*, of stimulating the other communities, its entropy captures how many distinct communities the community influences. We refer to the strategy of targeting communities by means of the product of their total impact and its entropy as **impact focus**.

IV. Experiments

The main purpose of our experiments was to investigate information cascades with respect to three factors:

1) Number of targeted communities (q).
2) Number of users sampled from each targeted community for initial activation (s).
3) The capability of different heuristics to *predict* which communities to target in future such that as many nodes and communities as possible are active at the end of the spreading process.

In order to take the time into account, we considered *pairs* of consecutive snapshots of the reply-to network. Each snapshot was one time slice long: one week for Boards, two months for SAP. Using our three targeting strategies (given below), we selected target communities from the first snapshot, and then simulated the diffusion on the second snapshot using the two diffusion models (see Section III-C). This simulates a scenario of a stakeholder who uses knowledge of the current state of the system to *select* certain communities and then attempts to spread information through the system by posting into them. Since the seed actors were sampled from the targeted communities, we repeated the simulation l times for different samples. Thus, we considered the mean value of the number of activated nodes at the end of each simulation for comparison. The simulation ended when it converged or when the maximum number of iterations, 500, has been reached. In fact, we observed that the diffusion process usually converged in ≈ 20 iterations.

In total, we evaluated three targeting strategies:

(a) **Impact focus** (IF) targets communities highly influencing many other communities (see Section III-D).
(b) **Group in-degree** (GI) was considered as a reasonably well-established centrality measure of communities. It is defined as the number of replies the members of a community received from the non-members [6]. Intuitively, it measures how much the community *in total* stimulates other communities. We chose group in-degree, because it is a generalisation of node degree, which has been widely used as a heuristic for influence maximization when targeting individual actors [3], [7]. Please note that the group in-degree, however, was not originally motivated by the influence maximization problem and here it is used to represent an intuitive and simple heuristic only.
(c) **Random** (R) was used as a baseline, and simply means a uniformly random choice of the communities to be targeted. For each combination of the number of targeted communities q and sampled users s, we repeated the simulation for a different sample of targeted communities l times, and averaged the results. Random targeting, especially in combination with high number of initially activated users, may be viewed as a spam targeting strategy. Therefore, the point at which its information spread converges suggests that it may be the same point at which the stimulation may become inefficient because it starts to be ignored by the users. I.e. if a heuristic for targeting a certain number of communities and users is performing no better than randomly targeting the same number of communities and users, then it is likely to be a 'saturation' point of targeting communities and users.

Since some of the snapshots were relatively large (see Table I), we set the number of repetitions l to 30 for the sake of computational tractability. Namely for the random baseline every combination of the number of targeted communities and user sample is repeated l^2 times.[3] This is done because both the target communities *and* the users within the communities were sampled.

In order to investigate the information cascades with respect to the main three factors in the two data-sets using the two diffusion models, we conducted four different experiments named in Table II.

	CALTM	CAICM
Boards	Boards-LTM	Boards-ICM
SAP	SAP-LTM	SAP-ICM

TABLE II: Names of the four experiments

As outlined previously, we organised each experiment into $T-1$ pairs of consecutive snapshots, i.e. the targeted communities were chosen based on the activity in snapshot t, and the simulation was run on the information diffusion network in the following snapshot, $t+1$. In each experiment we measured the spread of information over the users and over the communities depending on the parameters q, s, and the targeting strategy. For this purpose we defined two quantities: *user activation fraction* u and *community activation fraction* c. First, let A to be a set of all users that have been activated during the simulation. The user activation fraction u is then defined as $u = |A|/n$, the fraction of all the users that have been activated during the diffusion process. The community activation fraction is defined as:

$$c = \frac{1}{k}\sum_{j=1}^{k}\left(\frac{\sum_{i\in C_j\cap A}\mathbf{M}_{ij}}{\sum_{i=1}^{n}\mathbf{M}_{ij}}\right), \tag{5}$$

where C_j is a set of users of j-th community. The numerator in the brackets sums up the memberships of the

[3]The computation of the random baseline for Boards for one snapshot took for parameters $q \in [1,5], s \in [1,50], l = 30$ approximately 7 hours using 2 Intel Xeon CPUs.

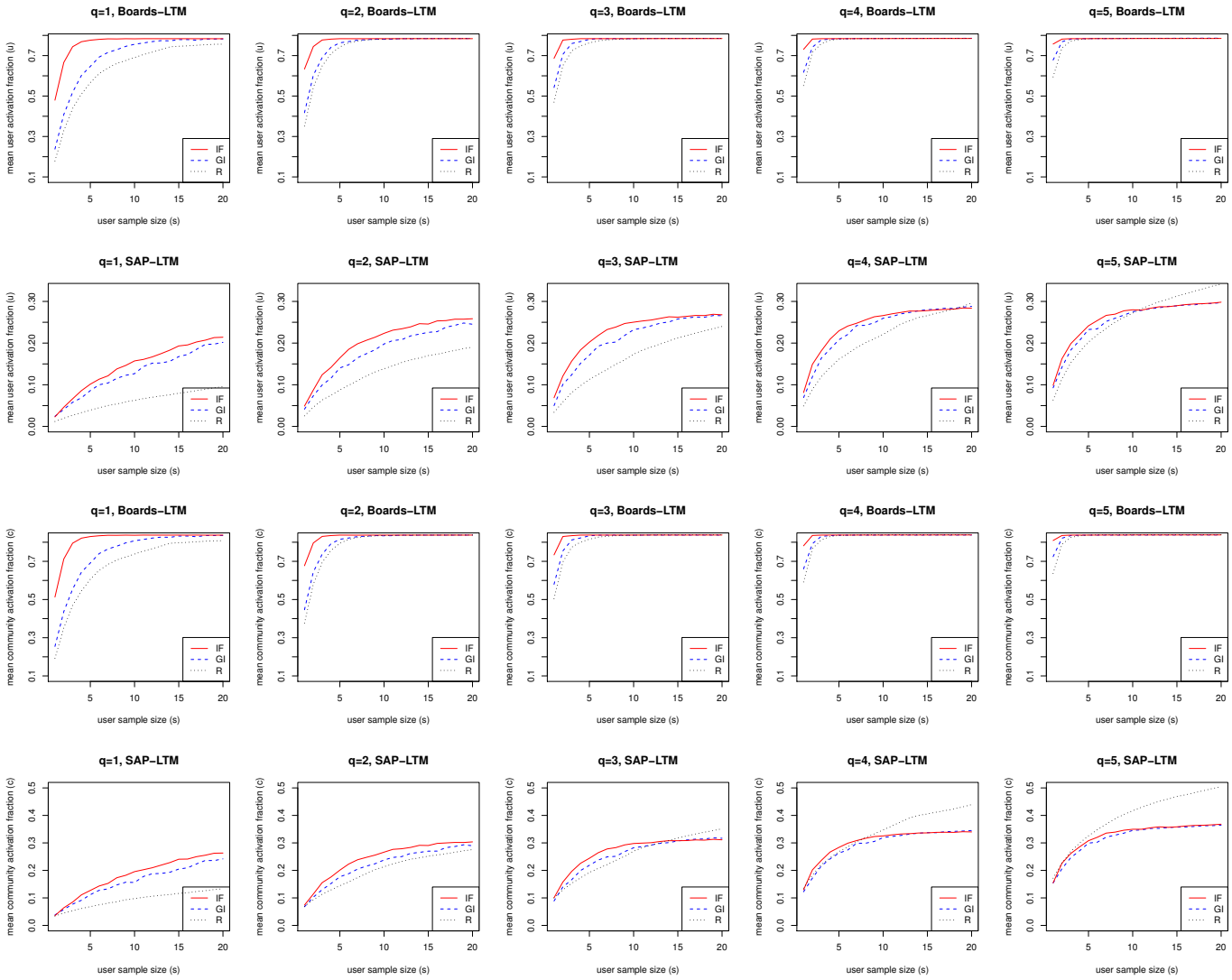

Fig. 2: Average user (u) and community (c) activation fractions obtained in Boards-LTM and SAP-LTM experiments. The plots for Boards-ICM and SAP-ICM experiments look similarly and we include them only in the supplementary material for space reasons. Please note that only values up to $s = 20$ are displayed for the sake of brevity.

activated members of the j-th community, and the divisor is the cardinality of the set representing the community. This represents a sum of fractions of community members that have been activated. This is then normalised by k, the total number of communities, of which is 1 if all the users in all communities were activated, and 0 if no users were activated. Please note that c respects the user activity patterns (by taking their memberships into account), and it also treats all the communities equally, which allows us to investigate diffusion across communities as opposed to individual users only.

V. Results

In this section we report on the results of the four experiments we conducted in order to investigate the information cascades across actors and communities. We used the two modified information diffusion models to simulate the cascades across the information flow network and we measured the performance of each targeting strategy by *user* and *community activation fractions*. First, we present initial insights into the average performance of the strategies. After that, we investigate rigorously which concrete combinations of factors induced one targeting strategy to have higher performance than the other two.

A. Average Performance of the Strategies

The first question we addressed was whether there is a difference between the *average* activation fractions achieved by the different targeting strategies. Therefore, for each targeting strategy at different values of q, we plotted the mean value of the activation fractions (u and c) over all the snapshots as a function of the number of sampled users (s). Figure 2 shows the plots for the Boards-LTM and SAP-LTM experiments. Since the plots for Boards-ICM and SAP-ICM experiments are similar, we include them only in the supplementary material. The upper two rows of Figure 2 show the values of user activation fractions u, while the bottom two rows contain the plots of community activation fractions c.

Overall, we found that in both data-sets the impact focus frequently achieved higher u and c than the other two strategies, especially for small $s \in [1, 20]$ and $q \in [1, 2]$. The higher

q	experiment	$u_{IF} > u_{GI}, u_R$	$u_{GI} > u_{IF}, u_R$	$u_R > u_{IF}, u_{GI}$	$c_{IF} > c_{GI}, c_R$	$c_R > c_{IF}, c_{GI}$
1	SAP-LTM	-	-	-	10	-
	SAP-ICM	-	-	-	11,12	-
	Boards-LTM	1–11	19,22,24,25,27,29,30,31,33–38,40,41	-	1–12,17	-
	Boards-ICM	1–12,15,20,23,32,39,47	-	-	1–16,18–50	-
2	SAP-LTM	6	-	-	6,9	-
	SAP-ICM	1,3,6,10,14,16	-	-	3,6,8,10,12–16	-
	Boards-LTM	1–5	12	-	1–5	-
	Boards-ICM	1–7	-	-	1–7,12–21,23–25,27–50	-
3	SAP-LTM	1,4,7–9	-	-	-	-
	SAP-ICM	1,3,6,7,11	-	-	6	-
	Boards-LTM	1–4	31,33,36–50	-	1–5	14–34,36–50
	Boards-ICM	1–5	48	-	1–5,8–10,12,14–50	-
4	SAP-LTM	2	-	34–50	-	13–50
	SAP-ICM	3,4	22	-	-	28,30–50
	Boards-LTM	1,2	24,27–50	-	1,2	9–50
	Boards-ICM	1–3	40–42,44,46,48–50	-	1–3,5,12,14–50	-
5	SAP-LTM	-	-	19–50	-	9–50
	SAP-ICM	3,4	-	-	-	1,10–50
	Boards-LTM	1,2	21–50	-	1,2	6–28,30–50
	Boards-ICM	1,2	17,19,26,28,29,31–50	-	1–4,13,15,16,18–50	-

TABLE III: **Number of sampled users** (s) for each number of targeted communities (q) when one strategy performed better than the other two with respect to either user (u) or community (c) activation fractions. For example, we see that for one targeted community ($q = 1$) in the SAP-LTM experiment IF performed better than GI and R, with respect to c for 10 sampled seed users ($s = 10$).

these factors were, the smaller the difference between the competing strategies. We also observed an effect of diminishing returns which suggests that by selectively targeting only one community, it is possible to efficiently penetrate a large part of the system, while the gain in spread from increasing the size of the target set became gradually smaller.

Furthermore, we see that for each number of targeted communities, the diffusion process became saturated at approximately 75% of users or communities activated (on average) in Boards. However, these figures were much smaller in SAP for impact focus and group indegree as these strategies usually achieved spread over approximately 30% of users or communities. These findings, along with the facts that there are much less edges per node in SAP than Boards (see Table I) and that it is much less common for a user to be active in multiple fora (see Section III-A), suggest that it is harder for a piece of information to diffuse in SAP as there is simply less interaction between its users.

The difference between the strategies was also generally smaller in SAP then in Boards. While the random strategy often performed worse in both data-sets, we observed that for high target sizes, especially for $q \in [4, 5]$, the random strategy outperformed the other two in SAP. We believe that this is caused by the nature of the information flow network in SAP, which we discuss at the end of this section.

B. When Does One Strategy Outperform the Other Two?

We observed that the average user and community activation fractions differed with respect to the targeting strategy, number of targeted communities q and number of users sampled from the communities s. We now want to investigate these differences rigorously by means of hypotheses testing. Each of the hypotheses state that one targeting strategy, e.g. IF, performed better than the other two, e.g. GI and R, with respect to either user activation fraction u or community activation fraction c. These are conveniently written as $u_{IF} > u_{GI}, u_R$, and $c_{IF} > c_{GI}, c_R$, respectively. As there are three such possible orderings and two performance measures, the total number of these *main* hypotheses is six.

In order to test each of the main hypotheses, we broke them down into individual *pair tests*, e.g. to test a hypothesis $u_{IF} > u_{GI}, u_R$ we first evaluated $u_{IF} > u_{GI}$ and then $u_{IF} > u_R$. Each of these individual tests was performed on a paired sample of values of a performance measure (u or c), that was obtained from the simulation runs. Because we cannot assume the network to remain the same over the time, we also cannot assume the values of the performance measure to be normally distributed. Therefore we used a non-parametric Wilcoxon signed-rank test [15]. Finally we accepted the main hypothesis, e.g. $u_{IF} > u_{GI}, u_R$, if the individual tests remained significant ($\alpha = 0.01$) after applying the Bonferroni correction.

Table III shows the values of parameter s (number of sampled users from the target communities) for which we accepted one of the main hypotheses. The table is divided into three parts. The left column lists the number of targeted communities q and the experiments in which the hypotheses were tested. The middle column lists the results for user activation fraction u, and the right column lists the community activation fraction c results. Please note that as the Wilcoxon test does *not* compare *average* values over time, the plots in Figure 2 may not wholly correspond to Table III.

From this table, we see that our initial observation that IF performs better than the other strategies with respect to **user activation fraction** for small numbers of sampled users is largely confirmed. For instance, for one targeted community ($q = 1$) we see that IF outperformed the other two up to $s = 11$ in Boards-LTM experiment, and up to $s = 12$ (as well as $15, 20, 23$, etc.) in Boards-ICM. A similar behaviour can also be seen for other values of q, but only for a smaller number of sampled users. Analogously, in SAP data-set IF performed

better then IF or R for small s and $q \in [2, 5]$. However, we also see that GI tended to achieve higher user activation fractions than the other two for higher s, especially in the Boards-LTM experiment. This suggests that if the information is likely to be *adopted* by only a small number of seed users, then IF is a more suitable strategy, whereas if it is likely to be adopted by large number of seed users, then GI is likely to achieve higher spreads over the users.

The first interesting observation regarding performance of the strategies with respect to **community activation fraction** c is that we found that there were no cases where GI performed better than the other two strategies, which was why we omitted the corresponding column in the table. We generally see that IF again achieved higher c than GI or R. Especially for $q \in [1, 2]$, IF outperformed the others for a range of s in all the experiments. Taking into account that the gains in performance with additional increase in q and s featured diminishing returns, it thus seems more efficient to target only few communities by IF.

However, for $q \in [4, 5]$ and higher s, we observed that R strategy achieved higher u than the other two in the SAP-LTM experiment and higher c in all experiments except Boards-ICM, for which IF remained the best. We believe that this was caused by the fact that the information flow networks were fragmented into multiple connected components. The investigation of the connected components revealed that there was one large connected component on average accounting for 63% of the users in SAP and 92% of the users in Boards. We also found that both the IF and GI were highly *biased* towards targeting communities within the largest connected component (LCC). As the random strategy outperformed the other two namely with respect to c, we investigated how many communities and to what extent they were isolated from other communities. We found that whereas 80% of the communities in Boards had at most 20% of their members unreachable from another community, i.e they were isolated within the community, in SAP 25% of the communities had at least 75% of their members isolated.

As IF and GI were biased towards targeting communities within the LCC, the isolated communities were likely to be missed. For a sufficiently high number of targeted communities (in our case $q \in [4, 5]$), the likelihood that both the communities within the LCC and the isolated ones are targeted, is higher for R than for IF or GI. Since the community activation fraction c treats all communities as equal, and because random targeting mostly led to lower u, the higher performance of R can be attributed to its coverage of small peripheral communities, which account only for a small fraction of the total user base. We report on our analysis of connected components in more detail in the supplementary material.

VI. Conclusion

The results have shown that our framework achieved high user (u) and community (c) activation fractions and that it outperformed the other two strategies in the majority of the experiments for small numbers of targeted communities q and sampled seed users s. While GI performed better than the other two with respect to u for high s, notably in Boards, it never outperformed IF nor R with respect to c. Therefore, our framework was the only efficient targeting strategy with respect to both performance measures.

However, for high q (namely $q \in [4, 5]$) and s we observed that random targeting performed better than the other two strategies with respect to the community activation fraction, especially in SAP data-set. The cause was likely due to the fact that the underlying information flow network was fragmented into connected components, which made it hard to reach peripheral communities. Therefore, if SAP stakeholders aim for a community-inclusive targeting strategy it may be more appropriate to target communities within components associated with a particular topic of their message. The extension of our framework with topical analysis is thus our intended future work.

Acknowledgements

Research presented in this paper was supported by Science Foundation Ireland under Grant No. 08/SRC/I1407 (Clique: Graph & Network Analysis Cluster) and Grant No. SFI/08/CE/I1380 (Lion-2), and by the EU under Grant No. 257859 (ROBUST). We would like to thank Adrian Mocan and Falk Brauer from SAP, and Marcel Karnstedt and Donn Morrison from DERI for providing many helpful suggestions and comments.

References

[1] M. Newman, "Spread of epidemic disease on networks," *Physical Review E*, vol. 66, no. 1, p. 016128, 2002.

[2] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Physical review letters*, vol. 86, no. 14, pp. 3200–3203, 2001.

[3] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. of the ACM SIGKDD*. ACM, 2003.

[4] V. Belák, S. Lam, and C. Hayes, "Cross-community influence in discussion fora," in *Proc. of the AAAI ICWSM*. AAAI, 2012.

[5] ——, "Targeting online communities to maximise information diffusion," in *Proc. of the Workshop on Mining Social Dynamics (MSND) WWW'12*. ACM, 2012.

[6] M. Everett and S. Borgatti, "The centrality of groups and classes," *J. of Mathematical Sociology*, vol. 23, no. 3, pp. 181–201, 1999.

[7] J. Sun and J. Tang, *Social Network Data Analytics*. Springer, 2011, ch. A survey of models and algorithms for social influence analysis, pp. 177–214.

[8] M. McGlohon and M. Hurst, "Community structure and information flow in USENET: Improving analysis with a thread ownership model," in *Proc. of the AAAI ICWSM*, 2009.

[9] H. Wu, J. Bu, C. Chen, C. Wang, G. Qiu, L. Zhang, and J. Shen, "Modeling dynamic multi-topic discussions in online forums," in *Proc. of the AAAI 2010*, 2010.

[10] X. Song, B. Tseng, C. Lin, and M. Sun, "Personalized recommendation driven by information flow," in *Proc. of the ACM SIGIR*. ACM, 2006, pp. 509–516.

[11] X. Shi, J. Zhu, R. Cai, and L. Zhang, "User grouping behavior in online forums," in *Proc. of the ACM SIGKDD*. ACM, 2009, pp. 777–786.

[12] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge U. P., 2009.

[13] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[14] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[15] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.