Targeting Online Communities to Maximise Information Diffusion

Václav Belák
DERI, NUI Galway
IDA Business Park
Lower Dangan
Galway, Ireland
vaclav.belak@deri.org

Samantha Lam
DERI, NUI Galway
IDA Business Park
Lower Dangan
Galway, Ireland
samantha.lam@deri.org

Conor Hayes
DERI, NUI Galway
IDA Business Park
Lower Dangan
Galway, Ireland
conor.hayes@deri.org

ABSTRACT

In recent years, many companies have started to utilise online social communities as a means of communicating with and targeting their employees and customers. Such online communities include discussion fora which are driven by the conversational activity of users. For example, users may respond to certain ideas as a result of the influence of their neighbours in the underlying social network. We analyse such influence to target communities *rather than individual actors* because information is usually shared with the community and *not* just with individual users. In this paper, we study information diffusion across communities and argue that some communities are more suitable for maximising spread than others. In order to achieve this, we develop a set of novel measures for *cross-community influence*, and show that it outperforms other targeting strategies on 51 weeks of data of the largest Irish online discussion system, Boards.ie.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: SOCIAL AND BEHAVIORAL SCIENCES—*Sociology*; H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Systems and Software—*Information networks*; I.6 [**SIMULATION AND MODELING**]: General

Keywords

information diffusion, cross-community dynamics, discussion fora, dynamic online communities

1. 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, e.g. to promote awareness of company policy or its products and services. 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 [9, 11] to the spread of influence through a social network [7].

In the case of discussion fora, the scenario is somewhat different to information or action 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.

- Our **main hypothesis** is that it is possible to efficiently *engage* or *stimulate* a substantial part of the system by selectively targeting specific communities.
- We define this engagement in terms of **impact**, which measures how likely the average user in a given community would *generate activity*. We measure this activity in terms of *replies*.
- The **main problem** is then formulated as a *prediction* of the set of communities to target such that the stimulation of the system is maximised in the future.

To the best of our knowledge, our hypothesis and main problem has not yet been addressed. We provide a framework to identify communities for this maximisation problem and show that overall, it achieves a better network spread than other investigated targeting strategies.

1.1 Intuition of Impact

We are particularly interested in finding fora which have *impact* on other fora, i.e. users from one forum, on average, stimulates another forum to have a high number of replies (i.e. activity). For example, Figure 1 shows two discussion communities, $A = \{a, b, c, d\}$ and $B = \{b, c, d, e, f, g\}$, in which the nodes represent users connected by their replies. The thickness of the links reflect the number of replies. A user is *devoted* to one forum if the majority of his posts are to that forum. A user's community affiliation is reflected in the shading, such that the darker the node, the more a user is devoted to forum A, and the lighter it is, the more it is devoted to forum B.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

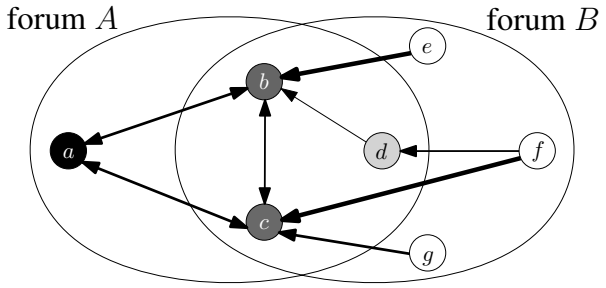


Figure 1: Example of impact from forum A to B. Nodes are users connected by links whose thickness reflects the number of replies. The shading expresses community affiliations, such that the darker the node is, the more it is devoted to forum A.

In this idealised scenario, we see that users $\{e, f, g\}$ in forum B reply frequently to users $\{b, c\}$, who are mostly devoted to A. And so, while more devoted members of forum A tend to converse amongst themselves, e.g. as $\{a, b, c\}$ do, they also receive a lot of replies from users of forum B. In short, users from B react often to users who are more devoted to A than to B. We further note that users $\{b, c\}$ are *central* users in that they receive many replies in general. So, not only are users $\{b, c\}$ more devoted to A, they are also the most central users of B. Our intuition is that even though community B has more members than A, A has a high impact on B because the most central users of B are in fact, more devoted to A.

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 [14], in this paper our definition of community influence is specifically tied to the reply-to activity.

1.2 Contributions

To define impact, we developed a novel framework for cross-community impact analysis, which is based on *purely structural* features, derived from a *dynamic reply-to graph*. Our framework is flexible and can be extended to exploit other features such as content-based ones. However, in this paper, we do not consider the content of the posts between individuals in communities. Using our framework, we identify the most influential communities using an adaptation of the well known notion of actor centrality. We compare the communities chosen under this framework with those identified as 'central' groups as defined by [4]. We then target combinations of the most central/influential groups according to these metrics and evaluate their spread on the largest Irish online discussion system Boards.ie[1] using an information diffusion model. For example, when we used our proposed measures to target one initial community, we found that it outperformed the other targeting strategies by up to a factor of 2 in overall network spread.

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

- Provide a framework for identifying influential *communities* in a social network.
- Motivate and present a Community-Aware Independent Cascade Model for study of information cascades in the context of discussion communities.
- Evaluate the spread of information using the modified model on the communities chosen by our framework against communities chosen by other baseline metrics on 51 weeks of data from Boards.ie.

The remainder of the paper is organised as follows. In the next section we refer to the related work. The framework itself together with the data-set, its preparation, and the diffusion model used for the evaluation are discussed in Section 3. The experimental setup is clarified in Section 4 and the results of the experiments are then presented in Section 5. The last section further discusses the results, outlines potential applications and extensions of the framework, and summarises our intended future work.

2. 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 briefly discussed.

2.1 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 fora — 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 Boards.ie, its users can and do post in multiple fora which then receive replies from members of other fora.

Wu et al. [16] 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 also adopt this notion of information flow, which is also similar to how Song et al. [13] 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 [12].

The problem of finding influential actors within a social network has been intensively studied in social network analysis [15], although not so much on the level of communities. For the individual actors, a classic approach is to use a centrality measure like actors' degree. Everett and Borgatti [4] 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

[1]See http://www.boards.ie.

group members". Hence the group degree captures 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 simply aggregating the users' degree into one actor.

In our previous paper [1] we presented an extensible framework for cross-community influence. In the following, we briefly summarise the main concepts of the framework and based on them we develop a technique for identifying influential communities, through which the underlaying social network can be efficiently stimulated.

2.2 Diffusion Models

Several models of how information or an action diffuses over a social network has been proposed (see e.g. [14] for a recent survey). A problem of maximising the spread of information or influence was introduced first by Kempe et al. [7], who also generalised many previously defined models by the *Independent Cascade Model* (ICM). We use this model in our analysis as a starting point to gain initial insights into information cascades in the context of discussion communities. Some possible extensions are further discussed in Section 6.

The model considers a social network represented by a directed weighted graph $G = (V, E)$, where vertices V are the individual actors and a weighted edge $w_{i,j} \in E \subseteq V \times V, \forall j \in V : \sum_{i \in V \wedge i \neq j} w_{ij} \leq 1$ expresses a probability of an actor j to adopt a piece of information or an action from i. Each actor can be either *active* or *inactive* and the simulation proceeds stochastically in discrete steps where the activation spreads from the active nodes to the non-active ones as follows. The diffusion process starts 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 neighbors 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 neighbors succeeds, it becomes active in the next iteration $t+1$. The process stops when it converges or when the maximum number of iterations has been reached.

3. PRELIMINARIES

This section presents the framework we have developed for the measurement and exploration of mutual impact of communities and the data we used to evaluated it. First we describe Boards.ie, the data-set and system we analysed. Next, we describe how we derived the information flow network we use for evaluation from the reply-to network, as well as the diffusion model itself. Finally, we formally define the notion of cross-community impact and other related measures. We consider a general case of k *overlapping fuzzy communities* [6] and n users.

3.1 Boards.ie

Boards.ie is 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 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 fora 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 1 presents some basic statistics of the analysed data.

number of snapshots (t)	51
number of communities (k)	540
mean number of nodes per snapshot	5,298
mean number of edges per snapshot	26,484

Table 1: Elementary statistics about the analysed data-set.

Our problem is to predict which communities to target in the future based on past/current observations. However, rather than aggregating the data up until 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 resulting in a sequence $s_1, \ldots, s_t$.

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 fora, 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 out a suitable time-window size, we sampled 10,000 posts and investigated the distribution of $\tau(\cdot)$. We found that in approximately 84% of the cases a user has posted into another fora within 7 days, while 14 days 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. In order to investigate how different targeting strategies affect the diffusion process, we took the last 51 weeks of the 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.

3.2 Inferring the Information Flow Network

We derive our information flow network from the reply-to network of Boards.ie in a similar manner to [16] such that if j replies to i then there is 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:

$$w_{ij} = \frac{r_{ji}}{\sum_{l=1}^{n} r_{jl}} \quad (1)$$

Figure 2 shows how the flow is reversed when an information flow graph is derived the from the reply-to graph.

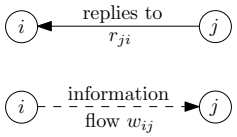
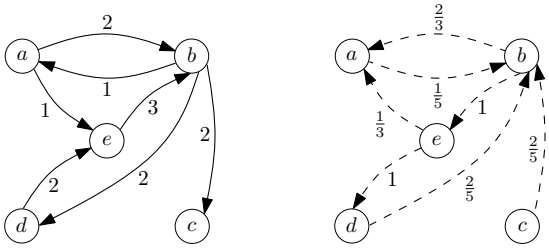


Figure 2: If j replies to i, it means that information has flowed from i to j. r_{ji} is the number of replies from j to i, w_{ij} is the weight of information flow from i to j.

Figure 3 gives a concrete example of how a weighted information flow graph is derived from a reply-to graph. Figure 3a shows the reply-to network, where the number of replies is the weight of the closest edge, e.g. there are two replies from user a to b. Figure 3b shows the induced information flow network, where the weights are defined by Equation 1, e.g. the information flow from b to a is the number of replies from a to b, 2, divided by the total number of replies out of a, 3.



(a) Original reply-to network

(b) Information flow network

Figure 3: An example reply-to graph and its induced influence graphs.

We note that there are alternative ways of inducing this network such as taking into account the authority of the node, e.g. defining w_{ij} as a proportion of in- versus out-degree, which we discuss in Section 6. However, we chose this normalising scheme as it is intuitive as a starting point.

3.3 Diffusion Starting From Communities

In the Independent Cascade Model (ICM) described in Section 2.2 the diffusion process starts from the set of seed nodes and *not* communities. Therefore, it is necessary to modify the model such that the process starts from a set of q targeted *communities*. Since the communities group together multiple actors and communities frequently share users in discussion fora, i.e. they overlap, there are multiple possibilities of how to extend ICM to the community level.

For instance, if we stimulate a certain community, e.g. by posting into it, should we assume that *all* users in this community become activated and then the spreading process unfolds, or rather that only a *subset* of the community's members becomes activated? If we assume all users are activated, it positively biases large communities over smaller ones regardless of the authority or participation patterns of their members. Moreover, this contrasts with the intuition that in a big community only a fraction of it is activated, because there is higher likelihood that some stimuli would be missed by some members — e.g. those ones who only occasionally participate in the community.

Therefore, since we do not know the likelihood that a user would respond to a stimulus, 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. Since the communities may overlap, the samples from distinct targeted communities may overlap as well. This corresponds to a scenario that the same user may be stimulated in different communities.

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. If the community size was smaller than s, we took all its users.

The modified *Community-Aware Independent Cascade Model* (CAICM) therefore proceeds as follows:

1. Select set T of q targeted communities.
2. For each community $C \in T$, 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 Model with L as a set of seed nodes.

The main parameters of CAICM are therefore the number of targeted communities q and the number of users sampled from each targeted communities s. In our experiments, we investigated 1–5 targeted communities, i.e. $q \in [1, 5]$, and $s \in [1, 50]$ users sampled from each targeted community.

3.4 Mutual Influence of Fuzzy Communities

We want to characterise to what extent one community is influencing another one as depicted in the ideal case in Figure 1. In that scenario, users mostly devoted to B, $\{d, e, f, g\}$, reply mainly to its central users $\{b, c\}$ who are mostly devoted to A, and therefore A has an impact on B. Thus, 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. In this section we show how to express and combine these 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, determined by a community detection algorithm [5], or from activity traces of the users. In our analysis we defined $m_{ij} = \frac{p_{ij}}{\sum_{l=1}^{k} p_{il}}$. Hence we measure the level of devotedness of a user by its activity in a similar manner to the work of McGlohon and Hurst [8].

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 [7] 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 [10], 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 [17] of the sets representing the communities — the sum of the columns of the membership matrix — in order to obtain a **normalised impact** matrix:

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

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 to 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 influence can be characterised as an entropy of the respective row of $\hat{\mathbf{J}}$. Because some elements of $\hat{\mathbf{J}}$ may be 0, let us first define a function $\rho(i, \hat{\mathbf{J}}) = \{l : l \in [1, k] \wedge \hat{\mathbf{J}}_{il} > 0\}$, that returns a vector of column indices of non-zero elements of i-th row of $\hat{\mathbf{J}}$. It is further necessary to normalise the rows of the matrix in order to obtain probability distributions of impact, i.e. $\hat{\mathbf{J}}^N_{i,j} = \hat{\mathbf{J}}_{i,j} / \sum_{l=1}^{k} \hat{\mathbf{J}}_{i,l}$. The normalised **impact entropy** of i-th community is then defined as

$$\mathcal{H}_I(i, \hat{\mathbf{J}}) = -\frac{\sum_{m \in \rho(i,\hat{\mathbf{J}})} \hat{\mathbf{J}}^N_{im} \log_2 \hat{\mathbf{J}}^N_{im}}{\log_2 |\rho(i, \hat{\mathbf{J}})|} \tag{4}$$

The entropy has range within $[0, 1]$. 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 in such case it 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 is one community 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**.

4. EXPERIMENTAL SETUP

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 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 week long. 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 CAICM (see Section 3.3). 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 the targeted communities. 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 3.4).

(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 [4]. 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 [7, 14]. 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 targetting strategy. Therefore, the point at which its information spread converges suggests that it may be the

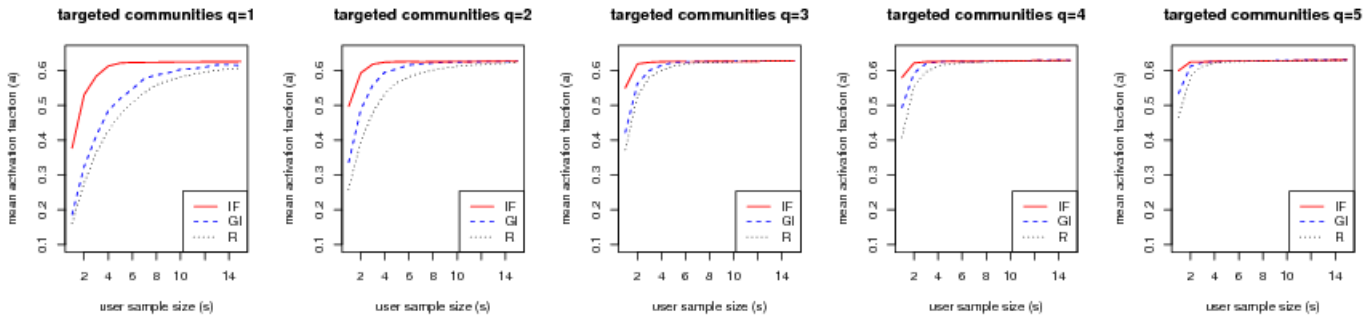


Figure 4: Average activation fraction (a) for different sizes of user samples s and number of targeted communities q. Since the saturation point was reached at approximately $s = 10$, only the values up to $s = 15$ are displayed for the sake of brevity.

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 1), 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.[2] This is done because both the target communities *and* the users within the communities were sampled.

As outlined in the previous paragraphs, we organised the experiment into 50 pairs of consecutive weeks, i.e. the targeted communities were chosen based on the activity in week t, and the simulation was run on the snapshot of the information diffusion network in the following week, $t + 1$. For each week, we thus obtained three values of the number of the activated users at the end of the diffusion process — one for each targeting strategy. A number of activated users at the end of the simulation was further normalised by the total number of users at each week $t + 1$. This *activation fraction*, a, thus represents the fraction of all the users that have been activated during the diffusion process.

Since our main interest was in the differences of activation fractions achieved by impact focus and group in-degree, we analysed three *types* of paired samples: (IF,GI), (IF,R), and (GI,R). Each of the types represents a group of paired samples obtained as follows. For each combination of q and s two activation fractions were computed for two targeting strategies, e.g. (IF,GI), in each week. Since the simulation was run on 50 weeks, the size of each paired sample was 100 (e.g. 50 activation fractions for IF and 50 for GI). In total, we had 250 ($s \times q$) samples for each type (IF,GI), (IF,R), and (GI,R).

5. RESULTS

In this section we report on the experiments we conducted in order to evaluate the three different strategies for targeting the communities. We used the Community-Aware Independent Cascade Model to simulate the information cascades over the information flow network. First, we show that the activation fractions achieved by each targeting strategy differed. Then we investigate which combination of factors induced one targeting strategy to have higher diffusion performance than the others.

5.1 Overall Maximum Spread Differences

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 targetting strategy at different values (q), we plotted the mean value of the activation fractions (a) as a function of the number of sampled users (s) (See Figure 4). We found that for $q = 1$ and $s \in [1, 10]$ the impact focus achieved a higher activation fraction than the other two strategies. However, we found that the higher q and s were, the smaller the difference between the activation fractions. Furthermore, we see that for each number of targeted communities, the diffusion process became saturated at approximately 60% of users activated (on average). This observation of *diminishing returns* 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.

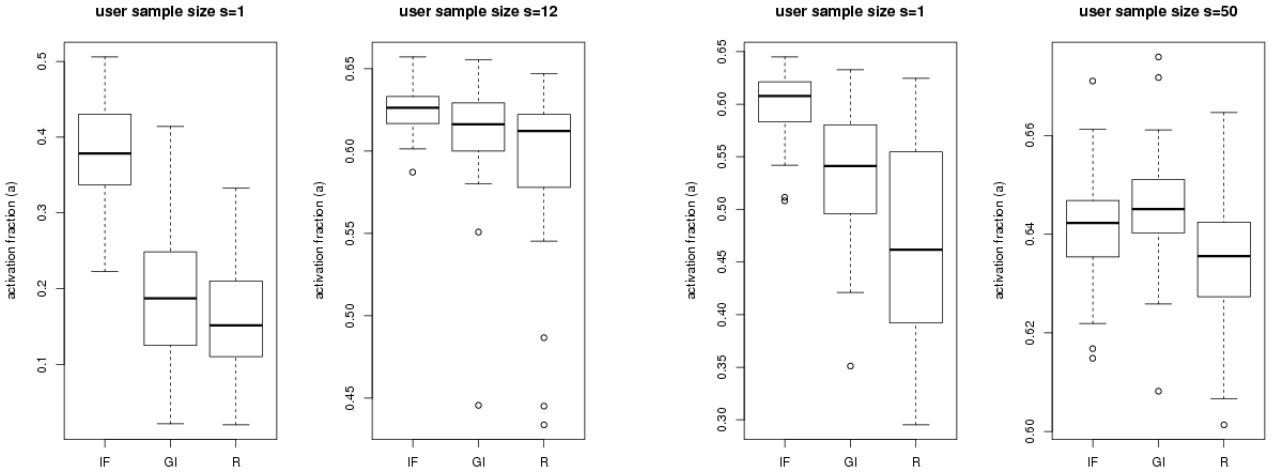
To confirm the hypothesis that there indeed is a difference between the strategies, we performed a three-way ANOVA to investigate the interplay between all three main factors of the experiment. According to the results we accepted the alternative that there was a significant difference between the average activation fractions achieved by each strategy (p-value 2.2×10^{-16}).

5.2 Does IF Achieve Better Spread than GI, R?

The second natural step was to find *which* combinations of the factors caused the difference. In particular, we hypothesised that the impact focus (IF) lead to activation of a larger fraction of the system in comparison to both the group in-degree (GI) and random baseline (R). First, we tested on each of the 250 samples whether IF achieved higher activation fractions than GI:

H_0 (null): The median difference of activation fractions achieved using impact focus (IF) and group in-degree (GI) was *lower or equal* than 0, in short $IF \leq GI$.

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



(a) One targeted community, and 1 and 12 users sampled from it.

(b) Five targeted communities, and 1 and 50 sampled users from each.

Figure 5: Activation fraction (a) for one and five targeted communities (q), and selected user sample sizes (s).

$\mathbf{H_1}$ (alternative): The median difference of activation fractions achieved using impact focus (IF) and group in-degree (GI) was *strictly higher* than 0, i.e. $IF > GI$.

Second, we tested a similar hypothesis with the alternative that impact focus lead to higher activation fractions than the random baseline, i.e. $IF > R$. Third, we evaluated the final hypothesis that the impact focus lead to higher activation fractions than the other strategies, i.e. $IF > GI, R$. We tested the hypotheses by the Wilcoxon signed rank test and adjusted the obtained p-values using the Bonferroni correction. Analogously, we also performed another set of tests with alternative hypotheses that the group in-degree led to higher activation fractions than the impact focus, and random baseline, i.e. $GI > IF, R$.

Table 2 lists the cases when the alternative hypothesis was accepted under the significance level $\alpha = 0.01$. We see that our initial observation, that the impact focus achieved higher activation fractions than the other strategies for smaller target sizes and user sample sizes, was confirmed. Namely for $q = 1$ the impact focus lead to significantly higher activation fractions than the group in-degree up to $s = 12$. As seen in Figure 5a, the median fraction of activated users for $q = 1$ and $s = 1$ was twice as high for the impact focus (0.38) than for the group in-degree (0.19). That means that in the strictest scenario where only one community was stimulated and only one user responded to that stimulation, a substantial part of the system was still penetrated — as opposed to the case when the group in-degree was used. It is also apparent that this difference was getting smaller with the rising s until it became non-significant, as depicted by the boxplot of the last significant case for user sample size $s = 12$ in the right part of Figure 5a.

The results were similar for all other target sizes except for the target size $q = 5$. In that case for smaller s impact focus again led to higher activation fractions than the group in-degree and the random baseline, but for higher s the group in-degree achieved higher a than the impact focus. This is further illustrated by Figure 5b, from which it can be seen that for the case $q = 5, s = 1$ the impact focus lead to significantly higher activation fractions than the group in-degree. On the other hand, the group in-degree led to higher activation of the actors for $q = 5$ and very high s. However, we also see that the activation fractions achieved by the impact focus were more stable for $s = 1$ as the corresponding standard deviation is lower for the IF (0.031) than GI (0.057) or random baseline (0.096). Finally, the difference in activation fractions for higher s is again relatively low — the median activation fraction for $s = 50$ was 0.642 for impact focus while for group in-degree and random baseline it was 0.645 and 0.636 respectively.

target size (q)	$IF > GI, R$	$GI > IF, R$
1	1–9, 12	-
2	1–5	-
3	1–3	-
4	1–2	-
5	1–2	35–37, 39, 45, 47–50

Table 2: Number of sampled users (s) for each number of targeted communities (q) for which either impact focus (IF) led to higher activation fractions than group in-degree (GI) and random baseline, i.e. $IF > GI, R$, or conversely for which $GI > IF, R$.

6. DISCUSSION

The results have shown that by the careful choice of the strategy it is possible to *efficiently predict* which communities to target in the close future such that a substantial part of the system becomes stimulated. In particular, we showed that the introduced measures of cross-community impact and its entropy led to higher spreads of stimulated users than the intuitive approach of targeting communities

by the total number of replies their members received from the rest of the system — the group in-degree. This was especially true for small numbers of targeted communities and small number of users initially activated in those communities. Even though this paper presents first insights and progress in the novel problem of targeting communities for maximising information diffusion, there are multiple interesting motifs for future research and in the rest of this section we highlight the most important ones.

Since we do not know how many users become initially activated, it remains an open question as to which concrete combination of factors are realistic for a given system. An interesting perspective is to estimate these parameters *directly* from the previous actions of the users where such information is available [2]. However, in the omnipresence of information overload, it is natural to assume that the more a chosen targeting strategy is efficient for small initial adoption likelihoods (while it remains stable for the higher probabilities), the more the strategy is a good heuristic whenever the exact information about the adoption probabilities is lacking. The impact focus proved to be particularly suitable for such cases.

The relatively fast convergence of the used model could be caused by the way the weights in the information flow network were obtained. While the presented weights are based on the likelihood of an information flow from one user to another, they do not take into account the node's authority. It is possible to hypothesise that the higher authority the node has (measured e.g. by its in-degree), the less likely it is to respond to an incoming stimulus. What is the best way to model the information flow network thus remains an important open question and again one interesting perspective is to *directly* exploit the past activity traces of the users [2].

Although communities are seen as a natural barrier for information diffusion [3][p. 506], we have observed that high spreads of information can be achieved by selectively targeting communities whose members are influential within many other communities. However, we also observed *diminishing returns* with targeting additional communities, which suggests that the targeted communities chosen by the impact focus, and group in-degree, were highly overlapping. This problem can be solved by greedy influence maximization approaches based on iterative estimates of spread gains, by means of Monte Carlo simulation [7]. However, this may become increasingly computationally intractable with a growing network size. An alternative and highly scalable approach to discriminate overlap between targeted communities is thus an avenue opened by the presented work.

Other interesting directions for future research opened by this paper are community-aware diffusion models of concrete, perhaps even conflicting, ideas or themes [2]. In the discussed cross-community scenario we assumed the information or stimulus is relevant to the system's user-base *in general*. However, in the case of when the aim is to address only the actors who are likely to be interested in *specific* information, such as a specific topic like sports say, it may be more efficient to amend the community targeting strategy adequately. Therefore, we aim to extend the presented framework with topic modelling, which in turn would enable investigation of how one community influences another one with respect to a given topic, and which communities to target in order to engage or stimulate only actors who are more likely to, in the future, respond to a given stimulus. Finally, in addition to the spread of information over the individual actors, we aim to investigate diffusion over *communities*.

7. ACKNOWLEDGMENTS

The material presented in this work is based upon works jointly supported by the Science Foundation Ireland under Grant No. 08/SRC/I1407 (Clique: Graph & Network Analysis Cluster) and under Grant No. SFI/08/CE/I1380 (Lion-2), and by the EU under Grant No. 257859 (ROBUST).

8. REFERENCES

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

[2] F. Bonchi. Influence propagation in social networks: A data mining perspective. *The IEEE Intelligent Informatics Bulletin*, 12(1), 2011.

[3] D. Easley and J. Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge U. P., 2010.

[4] M. Everett and S. Borgatti. The centrality of groups and classes. *J. of Mathematical Sociology*, 23(3):181–201, 1999.

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

[6] S. Gregory. Fuzzy overlapping communities in networks. *J. of Statistical Mechanics: Theory and Experiment*, 2011:P02017, 2011.

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

[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] M. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(1):016128, 2002.

[10] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[11] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200–3203, 2001.

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

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

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

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

[16] 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.

[17] L. Zadeh. A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with Applications*, 9(1):149–184, 1983.