# Community-Aware Content Diffusion: Embeddednes and Permeability

Letizia Milli[1,2] and Giulio Rossetti[1]

[1] KDD Lab. ISTI-CNR, via G. Moruzzi, 1, Pisa, Italy
{name.surname}@isti.cnr.it
[2] University of Pisa, Largo B. Pontecorvo, 2, Pisa, Italy
milli@di.unipi.it

**Abstract.** Viruses, opinions, ideas are different contents sharing a common trait: they need carriers embedded into a social context to spread. Modeling and approximating diffusive phenomena have always played an essential role in a varied range of applications from outbreak prevention to the analysis of meme and fake news. Classical approaches to such a task assume diffusion processes unfolding in a mean-field context, every actor being able to interact with all its peers. However, during the last decade, such an assumption has been progressively superseded by the availability of data modeling the real social network of individuals, thus producing a more reliable proxy for social interactions as spreading vehicles. In this work, following such a trend, we propose alternative ways of leveraging apriori knowledge on mesoscale network topology to design community-aware diffusion models with the aim of better approximate the spreading of content over complex and clustered social tissues.

**Keywords:** Diffusion, Epidemics, Community Discovery

## 1 Introduction

During the last decades, two topics above the others have been able to attract the interests of complex network analysis researchers continuously: community discovery [1] and epidemic/information spreading [2]. The former can be allegedly considered one of the hottest research tasks of this interdisciplinary playground. Every year, countless approaches are proposed to address such ill-posed problem, several papers that focus on the exploitation of network partitions for analytical purposes are published and, a few meta-studies underlying the limits of existing strategies for evaluating communities emerge. On the other side, epidemics, diffusion of information, gossip, and word of mouth phenomena represent central themes for a broad and heterogeneous research community. Computer scientists, as well as epidemiologists and physicists, have designed models to simulate content spreading on top of networked structures, often aiming at particular real-world scenarios (flu spreading, fake news, opinion dynamics to name a few).

When mixing the community discovery and the epidemics worlds a simple assumption is made, often by taking it as granted: the presence of a well-defined community structure accelerates the diffusive processes occurring among the nodes of the same group while, at the same time, acting as a “barrier” that delays such phenomena at the inter-community level.

Indeed, conceptually, both sides of such assumptions hold: the presence of a higher internal edge density of communities w.r.t. the number of edges that connect their nodes to the outside is a reason that, by itself, such behaviors are expected to appear. However, as already pointed out, there exist several community definitions as well as diffusion models with the latter rarely taking into account the former. Such a model rich playground naturally fosters an open question: *is it possible to define a more realistic family of diffusion models by explicitly leveraging the knowledge of node clusters presence?*

Indeed, in real-world scenarios, the diffusion of content might be affected by the polarization of the agents that spread it. An example is of such an effect is the diffusion of fake news in an online social network. Individuals in a social context tend to connect with like-minded peers, constructing cognitive filter bubbles (often enforced by the service as well) that act as a fertile ground for the diffusion of news (either fake or not) having a group-coherent content. Crossing the barriers of homogeneous node clusters such kind content might experience different degrees of spreading rate with respect to what they have within them. Taking into account such peculiarity is indeed a strategy worth considering while aiming at building a reliable proxy for content diffusion.

Starting from such an observation, in this paper, we introduce two community-aware diffusion models. The former model relates the node *embeddedness* within a community to its probability to foster a content to its neighbors; the latter, conversely, estimates such likelihood by assuming a user-defined degree of *permeability* of communities to contents coming from the outside. Both models move from the assumption above, implementing it with a different rationale. To understand the effects, such choices have on the diffusion speed and coverage, we then compare their simulations on a synthetic network scenario.

The paper is organized as follows. In Section 2, we introduce the two community-aware diffusion models, discuss their peculiarity and rationale. In Section 3, we compare the proposed models on synthetic networks with planted community structure. Finally, in Section 4 the literature relevant to our work is discussed and, Section 5 concludes the paper.

## 2 Community-Aware Diffusive Modeling

Diffusion processes are often modeled as context dependent phenomena. Different approaches have been proposed to simulate epidemics as well as opinion and content spreading, each one of them modeling specific peculiarities of the context they were designed to approximate. In our analysis we will focus on diffusive processes happening in social scenarios: for such a reason the models we design

**Algorithm 1** ICE

```
Require: $G = (V, E)$ a social graph; $I_0$ set of initial infected node
$I \leftarrow I_0$
$i \leftarrow 0$
while $|I| \leq |V|$ do
    $I_{i+1} \leftarrow \{\}$
    for $(u, v) \in E$ do
        if $v \in I_i \wedge u \notin I$ then
            if $\phi_{u,v} \neq \emptyset$ then
                $th(u, v) \leftarrow e_{u,v}$                    ▷ Embeddedness
            else
                $th(u, v) \leftarrow 1 - e_{u,v}$
            $p \leftarrow rand(0, 1)$                            ▷ Random value in [0,1]
            if $th(u, v) \geq p$ then
                $I_{i+1} \leftarrow I_i \cup u$
        $I \leftarrow I \cup I_i$
    yield $I_i$                                                  ▷ Return infected at time $i$
```

can be ascribed as variations of the well-known Independent Cascade (IC)[3] one[3].

**Independent Cascade Model.** The IC model, as most diffusion ones, takes as input (i) a social graph, (ii) a set of initially infected seed nodes $I_{t_0}$ and (iii) a diffusion probability $p_{u,v}$ – that, in principle, can be fixed independently for each node pair $(u, v)$. During a generic iteration of the IC simulation, each node can be either susceptible or infected. The simulation proceeds in discrete steps according to the following transition rule: during a generic iteration $t$ all nodes that have been infected at time $t - 1$ are considered active and had a chance to infect their susceptible neighborhood, with probability $p_{u,v}$. Independently from the success or failure of the infection, the infected nodes at time $t$ will not be allowed to spread the infection in consecutive rounds. The simulation runs until status transitions are no longer possible.

The IC model does not take into account the underlying network topology since it evaluates only local information (e.g., the presence of an edge among a pair of nodes and its associated activation probability) to implement the transition rule. To adapt it to our goal, we need to start analyzing the community/diffusion assumption. Such a postulate can be broken down into two statements: (i) community structure foster internal diffusion, (ii) community structure slow-down the diffusion across different communities. Indeed, depending on which of the two statements we decide to enforce in a community-aware diffusive model, we might observe different results. For such a reason, we design two alternative variants of the IC model, namely ICE (IC with Community Embeddedness) and ICP (IC with Community Permeability).

**ICE.** As a first way to embed community awareness into the IC model we designed an approach that ties the probability $p_{u,v}$ to the edge embeddedness within its community.

[3] Implementations of the proposed models are made available through the NDlib python library [4].

**Algorithm 2** ICP

```
Require: $G = (V, E)$ a social graph; $th$: the edge threshold;
$\eta$: the degree of permeability; $I_0$ set of initial infected node
$I \leftarrow I_0$
$i \leftarrow 0$
while $|I| \leq |V|$ do
    $I_{i+1} \leftarrow \{\}$
    for $(u, v) \in E$ do
        if $v \in I_i \wedge u \notin I$ then
            if $\Gamma(v) \cap \Gamma(u) \neq \emptyset$ then
                $th(u, v) \leftarrow th(u, v) * \eta$            ▷ Permeability
            $p \leftarrow rand(0, 1)$            ▷ Random value in [0,1]
            if $th(u, v) \geq p$ then
                $I_{i+1} \leftarrow I_i \cup u$
        $I \leftarrow I \cup I_i$
    yield $I_i$            ▷ Return snapshot status
```

**Definition 1 (Edge embeddedness).** *Given an edge $(u, v)$ with $u, v \in C$ its embeddedness is defined as:*

$$e_{u,v} = \frac{\phi_{u,v}}{|\Gamma(u) \cup \Gamma(v)|}$$

*where $\phi_{u,v}$ is the number of common neighbors of $u$ and $v$ within $C$ and $\Gamma$ compute the set of neighbors of a node in the analyzed graph.*

Algorithm 1 shows the pseudo-code for the ICE model. In this variation of the IC model, the values of the edge diffusion probability are not selected by the user but driven by the community partition considered. If $(u, v)$ is well-embedded within the community $C$ – and the community is adequately defined – the ICE model will tend to increase the diffusion probability significantly, conversely, if the edge is not well embedded (e.g., the nodes belongs to different communities, or at least one of them is peripheral to the community) it will reduce it.

**ICP.** Since ICE does not take explicitly into account the role of communities as "barriers" to content diffusion, we designed another ad-hoc model that leverages the concept of permeability. A community is "permeable" to content if it allows that content to spread from it quickly (or vice-versa, if it allows the content so easily be transmitted to the outside). Conversely, a community that dampens the diffusion probability across its border has a low degree of permeability.

Algorithm 2, shows the pseudo-code for the ICP model. The required parameters are the edges' threshold as well as the degree of community permeability, $\eta$ (that ranges in $[0, 1]$). At each iteration, for each edge $(u, v) \in E$, if $v$ is an "infected" node and $u$ is a "susceptible" one and they belong to the same community our method acts as a standard IC; instead, if the nodes belong to different communities, the probability $p_{u,v}$ is dampened of a factor $\eta$.

In the following section, we will test ICE and ICP both on synthetic networks having planted community structure to understand which are the effects of different modeling choices on the diffusion process unfolding in a controlled environment.

## 3 Experimental Analysis

In this section, we describe our experimental analysis, focusing our attention on three aspects: the network used, the experimental protocol adopted and the results obtained.

**Datasets.** One of the main issues related to our analysis lies in the ill-posedness of the community discovery problem. Our approaches need to know in advance the network decomposition in communities; however, plenty of community discovery algorithms has been designed so far, each one optimizing a different quality function and, as a consequence, producing different node partitions. Due to the absence of a *"one fits all"* approach, to effectively test our diffusive models we need to know in advance, for each potential network dataset, which community discovery algorithm is able to produce the highest-quality and optimally separated node clusters. Since it is likely that different datasets would require different algorithms, to make our analysis more reliable, we will focus only on synthetic generated networks having planted communities.

Tu such extent, we simulate ICE and ICP on networks generated with LFR[5], an algorithm that produce synthetic graphs having apriori known community structure . We generated nine different networks, each composed of 10000 nodes, having an average degree of 10, while varying the mixing parameter $\mu$ from 0.1 to 0.9 with a step of 0.1. Such a parameter allows us to specify how the nodes that belong to a community are connected to each other. Each node shares a fraction of $1-\mu$ of its links with the other nodes of its community and a fraction of $\mu$ with the other nodes of the network. Therefore, the threshold $\mu = 0.5$ marks the border beyond which communities are no longer defined in the strong sense, i.e., networks generated with $\mu < 0.5$ guarantee that each node has more neighbors in its community than in the others. In Table 1 are summarized the basic statistics of the generated networks and their communities.

**Experimental protocol.** Our experimental protocol, given a graph $G$, consists of the following four steps:

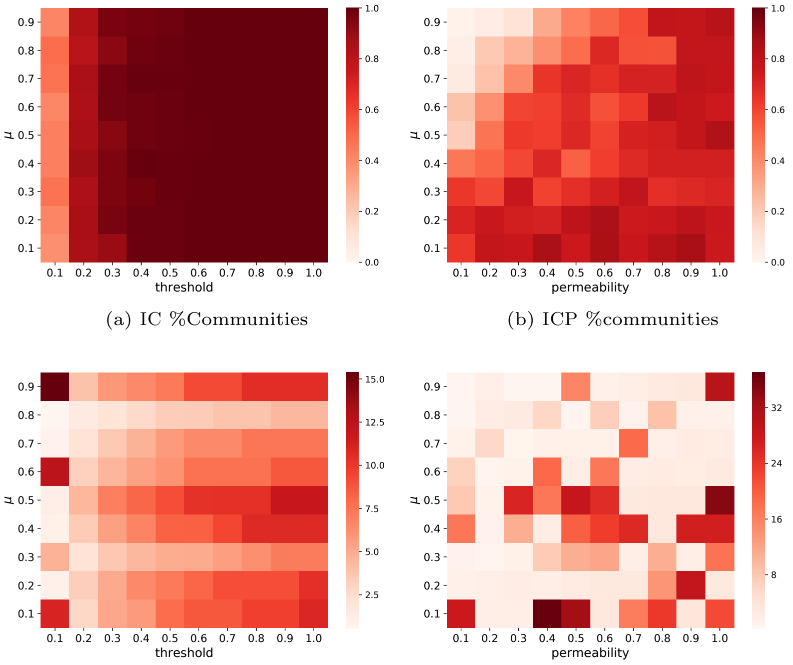
| Network | # Coms. | Avg Com. size | Std Com. size |
|---|---|---|---|
| 0.1 | 87 | 116 | 58 |
| 0.2 | 73 | 138 | 98 |
| 0.3 | 50 | 201 | 164 |
| 0.4 | 75 | 134 | 87 |
| 0.5 | 83 | 121 | 61 |
| 0.6 | 61 | 164 | 127 |
| 0.7 | 52 | 193 | 167 |
| 0.8 | 31 | 323 | 382 |
| 0.9 | 74 | 136 | 79 |

Table 1: LFR datasets statistics.

i. *Community identification.* Considering the LFR planted partition – that guarantee a complete coverage, non-overlapping, node clustering – we identify as $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$, with $|\mathcal{C}| = k$ the community set.
ii. *Infection seeds identification.* Our aim is to start the diffusion process from each community most central node. To such extent, for each community $C_i \in \mathcal{C}$ we identify the node with greater closeness centrality score (e.g., that node that minimizes the distances to all other nodes in its community).
iii. *Diffusion simulation.* We simulate three diffusion models, IC, ICE and ICP, as described in Section 2. Each model is instantiated $k$ times, seeding every execution from a different "infected" node from the list of nodes extracted (e.g., imposing each time a different community as the source of infection).
iv. *Evaluation.* Finally, we compare the obtained results by analyzing the ratio of communities reached at the end of each simulation and the average time to complete the execution for all the compared models averaged on the number of distinct simulations $k$.

**Analytical Results.** In this section, we reported the results obtained by the three methods; first of all we compare IC and ICP to show how the latter is able to slow-down/reduce the diffusion process across different communities. Then, we consider the ICE model and we discuss how, due to its definition, it allows to reach a major number of communities in minor time.

Figure 1(a-b) shows the heatmaps for the results obtained with IC, on the left, and ICP, on the right, for all the generated graphs. In the first row, every heatmap cell represents the percentage of the number of communities reached at the end of the execution for different LFR graphs (for each graph such value is obtained by averaging the simulations performed starting from different seeds). On the y-axis, we have the $\mu$ values considered, and in the x-axis the threshold (for IC) and the permeability values (for ICP). As we can notice, the ICP diffusion process infects a minor number of communities w.r.t. IC one. Such a behaviour is justified by the fact that ICP slows down the diffusion process while considering edges that crosses community borders.
Considering Figure 1(b) we can see how on the heatmap bottom rows, the percentage of infected communities is greater compared to the top, where the value of $\mu$ design communities that are less connected internally than externally. For small values of $\mu$ ($\mu < 0.5$), the communities are well defined; each node has more neighbors in its community than in the others. So we can see how row-wise, e.g. fixed a value of $\mu$, a clear trend emerges: the lower the permeability values the higher the "barrier" action performed by the communities. Moreover, we can notice that such phenomenon is heavily amplified for higher values of $\mu$, e.g. while considering poorly defined communities: indeed, in that scenario the number of "border" edges affected by the dampening is higher since community nodes experience a lower embeddedness (their community internal degree tends to be lower than the external one). Figure 1(c-d), shows the average time to complete the execution for the two models. Indeed, by looking at the heatmap scale, we can observe that the diffusion process is slower in ICP than in IC. We can conclude that ICP acts as expected: increasing the permeability parameter



(a) IC %Communities

(b) ICP %communities

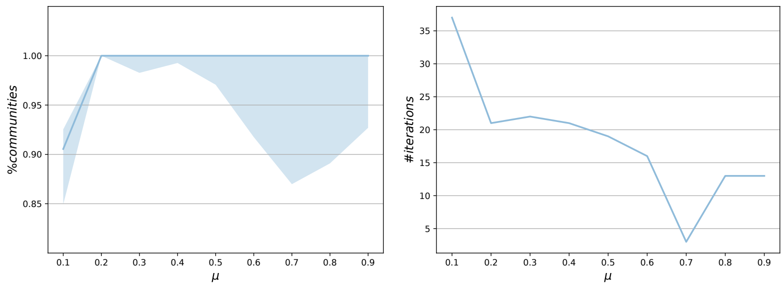
(c) IC #iterations

(d) ICP #iterations

Fig. 1: *Controlled topology scenario.* First row: heatmap of the percentage of communities reached at the end of the execution of the IC (left)and ICP (right). Second row: heatmap of the average iterations number required to stabilitze the simulation of the IC (left) and ICP (right). Heatmap rows identify different values of the LFR mixing coefficient ($\mu$) while columns the threshold values for IC and the permeability in ICP.

value, the diffusion process reaches a minor number of communities, and it takes a long time to converge to a stable state.

Figure 2(a-b), shows the results obtained by the ICE model. Being ICE parameter free – since the threshold value is computed as a function of edge embeddedness – we summarized the indicators using trend line plots identifying their point-wise average surrounded by interquartile ranges. In the Figure 2(a), the y-axis identifies the percentage of communities reached using ICE, the x-axis the LFR mixing parameter value. We can observe that ICE, that takes into consideration the structure of the communities to favor the diffusion among nodes belonging to a same cluster, allows reaching a highest percentage of communities than the other two models considered. In particular, while applying ICE on LFR graphs generated specifying $\mu \geq 0.2$ the process always converge to a complete community coverage. Conversely, in Figure 2(b) the y-axis identifies the number of iteration needed to reach the simulation stable state while the x-axis the pa-



(a) LFR %communities

(b) LFR #iterations

Fig. 2: ICE (a) Percentage of infected communities (areas identifies interquartile ranges) and (b) average number of iterations as function of $\mu$.

rameter $\mu$. From such plot we can notice that, as expected, the diffusion speed in ICE is inversely proportional to the value of $\mu$: the better defined the LFR communities are (lower values of $\mu$) the higher the number of iterations needed to reach convergence.

## 4 Related Works

To better contextualize our study three different, yet related, topics need to be reviewed: (i) diffusion process modeling, (ii) community discovery and (iii) the impact of community structure in diffusion dynamics networks.

**Diffusion Process modeling.** The notion of spreading or, contagion, underpins many widespread phenomena in biological, social, and technological systems. Since the early 20th century, researches have been trying to characterize and model diffusion processes systematically [6]. Mathematical models of transmission apply to a broad range of different phenomena; for example, the spread of information, innovations [7,8], and rumors can be modeled as a contagion process. With the increasing use of the online social network (OSN) as well as the pervasive use of mobile and wifi technologies in our daily life, we have access to a massive source of information. In recent years, researchers have developed a variety of techniques and models that have gained importance in the public health domain, especially in infectious disease epidemiology, by providing quantitative analyses in support of policy-making processes [9,10,11].

**Community discovery.** Community discovery the task of decomposing a complex network topology into meaningful node clusters is one of the hottest topics in complex network analysis [1]. Due to the extensive literature available in this area, over the years, several efforts were made to organize and cluster methods identifying some common grounds: due to the peculiar problem definition, thematic surveys emerged, focusing for instance on overlapping [12] and dynamic community discovery [13].

**Community structure and diffusion process.** Community structure has been shown to affect information diffusion, including global cascades [14,15], the speed of propagation [11], and the activity of individuals [16,17]. The presence of communities hinders epidemic spreading since this helps to confine the epidemics in the community of origin [11,18,19]. In [20], the authors study the impact of community feature on epidemic spreading, and the results show that the more communities a network has, the less the network is infected. In [21], the authors propose a set of local strategies for social distancing, based on community structure, that can be employed in the event of an epidemic to reduce the epidemic size. Recent empirical work suggested that modular structure may, counterintuitively, facilitate information diffusion [22]. Moreover, [23] investigates the impact of community structure on information diffusion with the linear threshold model. Other studies suggested that network modularity plays a more critical role in information diffusion than in epidemic spreading [24].

## 5 Conclusion

In this paper, we described two diffusion models that extend the IC one by introducing awareness of the network mesoscale community structure in the spreading process. ICE and ICP highlighted how enforcing different sides of the community/diffusion assumption led to a different impact on the resulting content diffusion speed in a controlled environment guaranteed by a pool of synthetic networks having planted community structure.

As future works, we plan to integrate community embeddedness and permeability in a single model to study the interplay of such characteristics. Moreover, we plan to apply the proposed models to a fake news real-case study by fitting their parameters with profiles extracted by users posting behaviors to understand the effect polarization and echo chambers have on content diffusion.

## Acknowledgements

This work is supported by the European Community's H2020 Program under the scheme "INFRAIA-1-2014-2015: Research Infrastructures", grant agreement #654024 *"SoBigData: Social Mining & Big Data Ecosystem"*[4].

## References

1. S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.
2. R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, "Epidemic processes in complex networks," *Reviews of modern physics*, vol. 87, no. 3, p. 925, 2015.

[4] SoBigData: http://www.sobigdata.eu

3. D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *9th ACM SIGKDD*, pp. 137–146, ACM, 2003.
4. G. Rossetti, L. Milli, S. Rinzivillo, A. Sîrbu, D. Pedreschi, and F. Giannotti, "Ndlib: a python library to model and analyze diffusion processes over complex networks," *International Journal of Data Science and Analytics*, vol. 5, no. 1, pp. 61–79, 2018.
5. A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Physical review E*, vol. 78, no. 4, 2008.
6. W. O. Kermack and A. McKendrick, "A Contribution to the Mathematical Theory of Epidemics," *Royal Society of London*, 1927.
7. F. Bass, "A new product growth for model consumer durables," *Management Sciences*, 1969.
8. E. M. Rogers, *Diffusion of innovations.* Simon and Schuster, 2010.
9. J. Leskovec, A. Singh, and J. Kleinberg, "Patterns of influence in a recommendation network," in *PAKDD*, pp. 380–389, Springer, 2006.
10. E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter," in *4th ACM WSDM*, pp. 65–74, ACM, 2011.
11. J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, "Structure and tie strengths in mobile communication networks," *National Academy of Sciences*, vol. 104, no. 18, pp. 7332–7336, 2007.
12. J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," *ACM Computing Surveys*, vol. 45, no. 4, p. 43, 2013.
13. G. Rossetti and R. Cazabet, "Community discovery in dynamic networks: A survey," *ACM Computing Surveys*, vol. 51, no. 2, p. 35, 2018.
14. A. Galstyan and P. Cohen, "Cascading dynamics in modular networks," *Physical Review E*, vol. 75, no. 3, p. 036109, 2007.
15. J. P. Gleeson, "Cascades on correlated and modular random networks," *Physical Review E*, vol. 77, no. 4, p. 046117, 2008.
16. M. S. Granovetter, "The strength of weak ties," *American Journal of Sociology*, 1973.
17. P. A. Grabowicz, J. J. Ramasco, E. Moro, J. M. Pujol, and V. M. Eguiluz, "Social features of online networks: The strength of intermediary ties in online social media," *PloS one*, vol. 7, no. 1, p. e29358, 2012.
18. X. Wu and Z. Liu, "How community structure influences epidemic spread in social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 2-3, pp. 623–630, 2008.
19. W. Huang and C. Li, "Epidemic spreading in scale-free networks with community structure," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2007, no. 01, p. P01014, 2007.
20. T. Liu, P. Li, Y. Chen, and J. Zhang, "Community size effects on epidemic spreading in multiplex social networks," *PloS one*, vol. 11, no. 3, p. e0152021, 2016.
21. Y. Bu, S. Gregory, and H. L. Mills, "Efficient local behavioral-change strategies to reduce the spread of epidemics in networks," *Physical Review E*, vol. 88, no. 4, p. 042801, 2013.
22. D. Centola, "The spread of behavior in an online social network experiment," *science*, vol. 329, no. 5996, pp. 1194–1197, 2010.
23. A. Nematzadeh, E. Ferrara, A. Flammini, and Y.-Y. Ahn, "Optimal network modularity for information diffusion," *Physical review letters*, vol. 113, no. 8, 2014.
24. L. Weng, F. Menczer, and Y.-Y. Ahn, "Predicting successful memes using network and community structure.," in *ICWSM*, 2014.