Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

With Flying Colors: Predicting Community Success in Large-scale Collaborative Campaigns

Abraham Israeli, Oren Tsur

Department of Software and Information System Engineering
Ben-Gurion University of the Negev, Israel
isabrah@post.bgu.ac.il, orentsur@bgu.ac.il

Abstract

Online communities develop unique characteristics, establish social norms, and exhibit distinct dynamics among their members. Activity in online communities often results in concrete "off-line" actions with a broad societal impact (e.g., political street protests and shifting norms related to sexual misconduct). While community dynamics, information diffusion, and online collaborations have been widely studied in the past two decades, quantitative studies that measure the effectiveness of online communities in promoting their agenda are scarce. In this work, we study the correspondence between the effectiveness of a community, measured by its success level in a competitive online campaign, and the underlying dynamics between its members. To this end, we define a novel task: predicting the success level of online communities in Reddit's `r/place` – a large-scale distributed experiment that required collaboration between community members. We consider an array of definitions for success level; each is geared toward different aspects of the collaborative achievement. We experiment with several hybrid models, combining various types of features. Our models significantly outperform all baseline models over all definitions of 'success level'. Analysis of the results and the factors that contribute to the success of coordinated campaigns can provide a better understanding of the resilience or the vulnerability of communities to online social threats such as election interference or anti-science trends. We make all data used for this study publicly available for further research.

1 Introduction

Communities, whether offline or online, play a crucial role in how we establish, perceive, and project our identity (Lewin 1947a; Zachary 1977; Ostrom 2000; McMillan and Chavis 1986; Côté 1996; Olson 2009). The fundamental role of the community, it's evolving norms, the dynamics between its members, its organizing principles, and collective action have been studied for decades, e.g., (Lewin 1947b; Granovetter 1973; Ostrom 2000; Fisher et al. 2019; Israeli, Kremiansky, and Tsur 2022), to mention just a few works.

The rise of online social platforms provides a unique opportunity to study phenomena that are associated with online communities organically and at a large scale (Melucci 1996; Lazer et al. 2009). It was shown that online activity often relates to or even inspires coordination off-line, such as support for a social change (Hässler et al. 2020), the financial markets (Lucchini et al. 2021; Mancini et al. 2022), street protest (Jackson and Foucault Welles 2016; Fisher et al. 2019), and violent outbursts (Peters et al. 2021).

The literature on large-scale studies of decentralized community operation and coordination is limited. Furthermore, research providing insights into the factors contributing to the successful execution of collective actions is scarce.

In this work, we aim to quantify, model, and predict the level of community success in a large-scale online campaign that requires collaboration among community members. Our definition of success deviates from traditional metrics such as the number of registered users or the retention rate of members within the community. Instead, we consider several measures, each capturing a slightly different aspect of the notion of success in a concrete campaign: accounting for the complexity of the campaign objective, the community resources, or the opposition it faces. Furthermore, our prediction models are interpretable, allowing us to analyze the contribution of different factors to the success level. This analysis, in turn, provides novel insights and validates the existing theory.

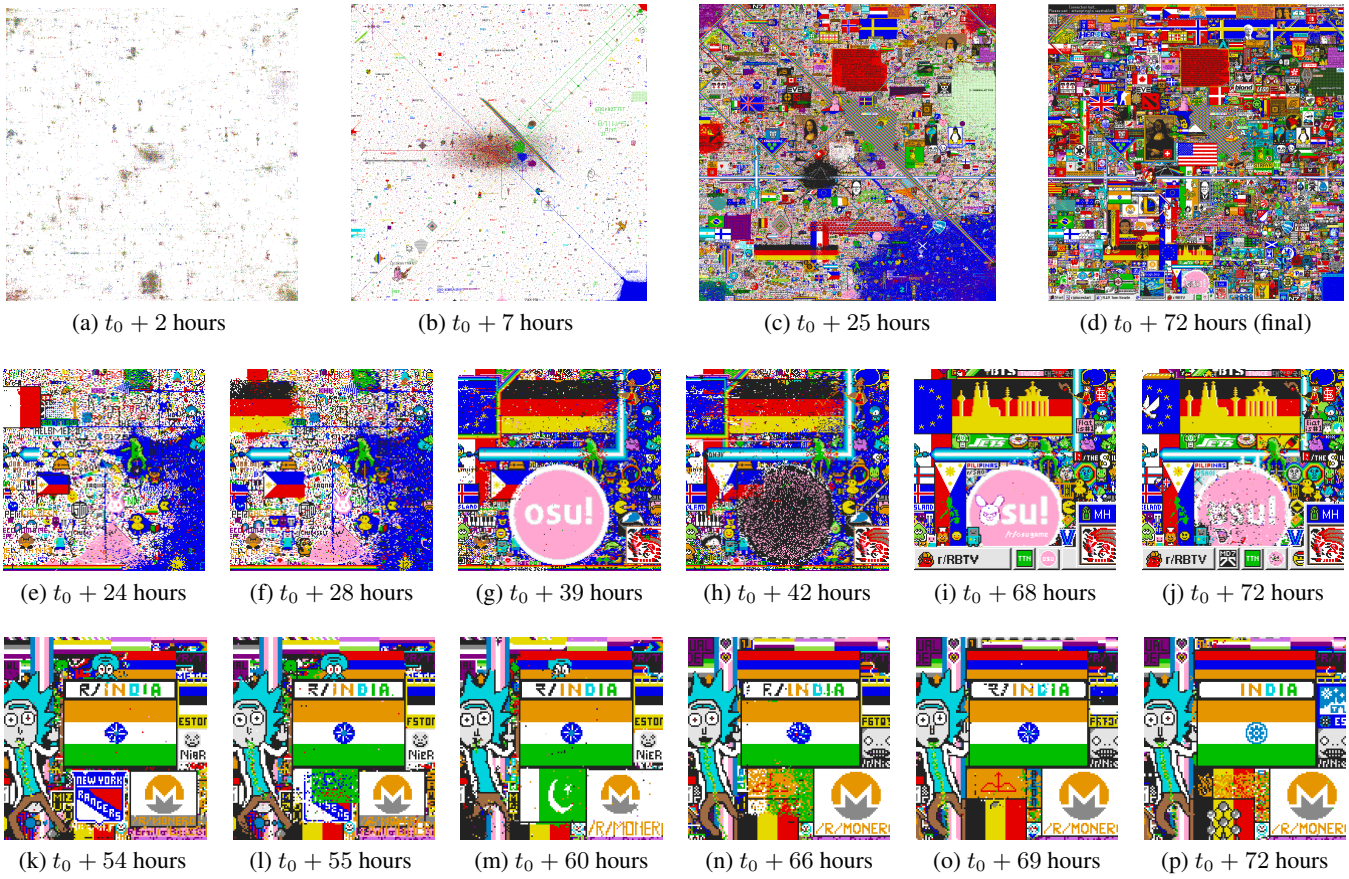
Reddit and the `r/place` experiment Reddit[1] is one of the most popular social platforms worldwide. On average, it attracts more than 430 million active users per month (Todorov 2022) that communicate at over 3.4M forums (as of December 2022), called *subreddits*. In each subreddit, users (*redditors*) can initiate a discussion thread, contribute to a thread, and up/down-vote other posts. Each subreddit constitutes a community that develops informal norms and formal rules.

The `r/place`[2] experiment was launched by Reddit on April Fools' Day, 2017. A shared white canvas of one million pixels (1000 x 1000) appeared in a new subreddit called r/place. Redditors could select any pixel and change its color. Every change was reflected on the *shared* canvas, thus viewed by all "participants".

Once a Redditor recolored a pixel, he/she was automatically blocked by the system for some random time (5–20 minutes), effectively preventing any single Redditor from

[1] Reddit website: https://www.reddit.com

[2] `r/place` subreddit: https://www.reddit.com/r/place/



(a) $t_0 + 2$ hours (b) $t_0 + 7$ hours (c) $t_0 + 25$ hours (d) $t_0 + 72$ hours (final)

(e) $t_0 + 24$ hours (f) $t_0 + 28$ hours (g) $t_0 + 39$ hours (h) $t_0 + 42$ hours (i) $t_0 + 68$ hours (j) $t_0 + 72$ hours

(k) $t_0 + 54$ hours (l) $t_0 + 55$ hours (m) $t_0 + 60$ hours (n) $t_0 + 66$ hours (o) $t_0 + 69$ hours (p) $t_0 + 72$ hours

Figure 1: Snapshots of the evolution of the full canvas, and zoom-in on illustrative conflicts. *Top row*: Evolution of the full canvas; *Middle row*: The logo of the OSU video game is generated by the OSU community, then repeatedly vandalized by TheBlackVoid community; *Bottom row*: The New York Rangers, Pakistan, Shiv Sena, and Belgium clash over the control of the area just under the Indian flag. The top row snapshots are taken from Israeli, Kremiansky, and Tsur (2022), and the other snapshots were generated from the archive gallery of the `r/place` provided by Albini (2017). A high-resolution image of the final state of the canvas is available at https://bit.ly/39e1E9a.

having any significant or lasting effect on the canvas. The `r/place` was not conceived with any specific purpose or goal, thus users were not encouraged to do anything in particular. To the users' surprise (and dismay?), the canvas was blocked for further manipulation after 72 hours.

Over the course of the experiment, the canvas was manipulated over 16 million times by 1.2 million Redditors. We refer to `r/place` as a naturally occurring large-scale controlled experiment. Figure 1 (top row) presents four snapshots attesting to the progression of the canvas' state, from its early chaotic state to its final shape – a diversified collage of complex logos, flags, symbols, and artworks.

Identity, clashes, and success in `r/place` In order to provide the appropriate definitions of success in a collaborative campaign, we should first present some of the dynamics that unfolded during the 72 hours of the experiment.

Examining the final state of the canvas (Figure 1d), one can observe that most of the artworks are associated with a well-defined identity, e.g., national flags, mascots, emblems of colleges and sport clubs, or gaming communities. Interestingly, a number of *new* communities were established during the experiment.[3] The most successful and recognizable ones are the monochromatic efforts – 'The-Blue-Corner' (bottom right in Figures 1b-1d), and 'The-Black-Void' (TBV) (middle of Figure 1c).

TBV was an organized trolling effort to vandalize the canvas by recruiting many Redditors to expand a black fractal-like shape in multiple regions, overriding other artworks. For example, consider the logo of the OSU video game, created gradually (Figures 1e – 1g), to be "raided" by TBV (see Figure 1h). The OSU logo is later recovered (Figure 1i) and attacked again, though with limited success (notice the black pixels scattered on the logo in 1j).

It is important to note that while the declared purpose of TBV's was pure vandalism, clashes between communities competing for "real estate" were common. An example is provided in Figures 1k – 1p in which The New York Rangers, Pakistan, Shiv Sena (Indian nationalist party), and

[3]Note that while these organizations are new, only pre-existing users could manipulate the canvas.

Belgium clash over the control of the area just under the Indian flag. Also note that while the Belgium community managed to successfully expand their flag, dominating the area at the termination of r/place, other communities were not as fortunate, having to relocate or disappear.

The r/place experiment provides a straightforward and (almost) unmoderated setting in which only coordinated efforts could have had any significant impact on the final state. As such, it provides a unique opportunity to study, on a large scale, how decentralized communities coordinate to achieve a common goal in a state of "emergency" in which their efforts are hindered and sabotaged by adversaries. Understanding the factors that contribute to the success of coordinated campaigns may shed new light (or validate social theory) regarding the resilience or vulnerability of communities to manipulation campaigns such as election interference or anti-science trends.

Measures of success While seemingly straightforward, success could be defined and measured in multiple ways. These measures may be correlated to a certain degree but are not identical. The simplest indication of success is binary: whether a community managed or failed to leave any recognizable mark on the canvas. Viewing success through this generic binary lens postulates that a community achieving the placement of a logo of a hundred pixels is as successful as a community achieving a logo of a thousand pixels. Another simple measure of success could be the number of pixels a community placed. However, ranking the success level based on pixel counts ignores many other factors that should be accounted for. Some relevant factors are the size of the community, the complexity of the logo, or the demand for the location of the logo on the canvas.

In this work, we explicitly consider these factors (size, location, complexity) and train prediction models to predict success according to the different measures. A further discussion, concrete examples, and formal definitions of *success* are provided in Section 2.

Our prediction models take into account a multifaceted representation of each community, combining linguistic patterns (e.g., vocabulary and distributional semantics), the community social structure (e.g., network embedding), and meta-features (e.g., number of active members and age). The full list of feature types is provided in Section 4.1, and the prediction models we consider are described in Section 4.2. We make all code, annotated data, models, and information for reproducibility of the research publicly available on the project's GitHub repository.[4]

Finally, we analyze the results and discuss the contribution of various input representations and community features to the success of a community, accounting for the different definitions of success. Moreover, we consider two modes of community representation: (i) Using data that was generated only *prior* to the experiment, and (ii) Representation based on data that was generated only *during* the 72 hours of r/place. These two modes allow us to further explore whether communities that quickly adapt to the "state of emergency" perform better. Results and analysis are presented in Section 5.

[4]https://github.com/NasLabBgu/rplace-predictions

2 Definitions of Success

As mentioned in the Introduction, the definition of community success (in a campaign) is not straightforward. Cunha et al. (2019) identified a number of ways to measure the success of a community (e.g., retention rate, growth of membership). However, these measures of success relate to the general state of online communities rather than to the successful coordination toward a specific goal or in mitigating a specific threat.

In this section, we discuss some contextual factors and propose five definitions by which success could be measured. These definitions would serve to assign labels (one binary and four continuous) to be predicted.

Leaving a mark (binary) The simplest indication of success is binary: whether a community managed or failed to leave any recognizable mark on the canvas by the end of the experiment. This concept of success is demonstrated in the bottom row in Figure 1, where the New York Rangers, Pakistan, and Shiv Sena failed to leave a mark as they were eventually overridden by the successful Belgium community.

This naive approach to success is problematic as it postulates that a small yet recognizable logo of a few pixels indicates the same success level of a community achieving a much larger logo. We, therefore, wish to consider a success measure that allows ranking, reflecting a success level.

Continuous measures of success Looking at the *number of pixels* a community has managed to place is a well-defined measure that allows ranking of the success levels. However, this crude measure forgoes many factors that should be considered, marking even smaller logos as a great success. Some of these factors are: *community size* – small communities are expected, a-priori, to place fewer pixels; The *complexity* of the campaign objective – logos with high entropy are harder to coordinate and maintain; *Shape* – longer "borders" are harder to protect (and clashes may erupt in multiple fronts); *Location* – some areas of the canvas are at a higher demand, thus harder to maintain and protect.

The impact of these factors on the definition of success is illustrated in Figure 2. The regression line (blue) indicates a positive correlation between the size of the community and the size of the logo (number of pixels placed). However, consider the following two gaming communities: r/osuGame and r/LeagueOfLegends (marked by the green arrows in the figure). Both communities allocated about the same number of pixels (6421 and 7114, respectively). However, the number of community members in r/LeagueOfLegends is more than an order of magnitude larger than that of r/osuGame (1.96M and 75.2K, respectively), suggesting that OSU is more successful. On the other hand, the entropy of the LeagueOfLegends logo is much higher than the entropy of OSU (indicated by the darker color), requiring more effort to create and maintain compared to the simplicity of OSU.

Finally, notice that the area of the OSU logo was in much higher demand compared to that of LeagueOfLegends (indicated by the size of the marker in Figure 2), suggesting

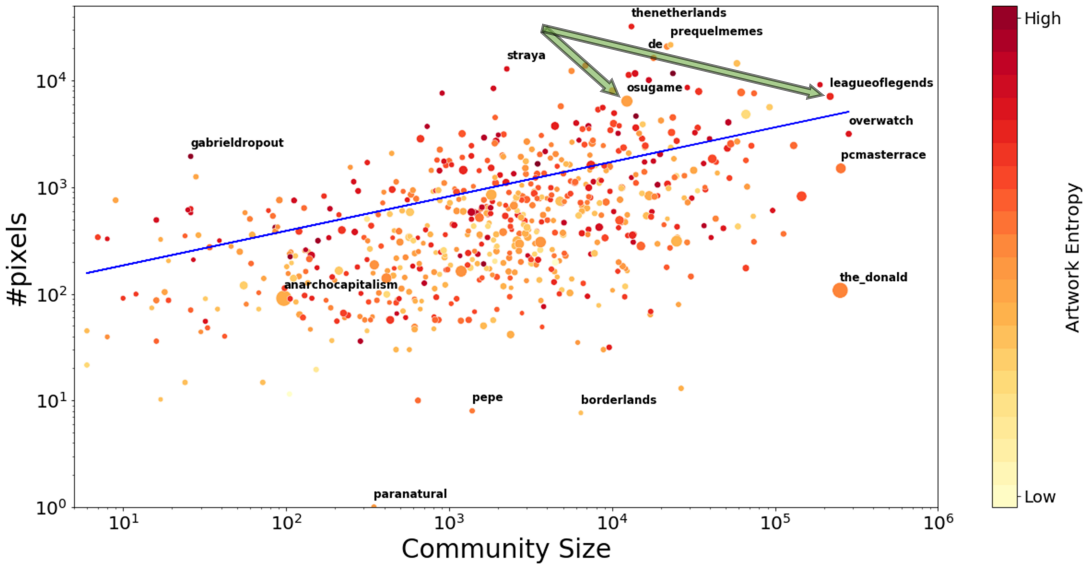


Figure 2: Success level vs. community size. Each point represents a participant community in the r/place experiment. Highlight communities are explicitly named. The size of each point represents how popular the region in which the community acted was. Colors emphasize the complexity of the drawn artwork, measured by the entropy. The blue line is a linear fit to the data. We use the log scale on both axes for better visibility of the figure.

that although similar in size, leaving a mark of this magnitude in the face of a fiercer opposition is a more impressive accomplishment.

We, therefore, define four additional measures of success, each provides a continuous success score that would be used as non-categorical labels – all are based on the number of pixels in a logo at the final state, factored by some function γ_k to account for different contexts. The success score of a community c^i with respect to k is therefore given by:

$$s_k(i) = \gamma_k(l^i_{\#pixels})$$

where, $l^i_{\#pixels}$ denotes the number of pixels in the logo l^i created by c^i, and $k \in \{\phi, |c|, p, d, H\}$ indicates the term by which $l^i_{\#pixels}$ should be factored:

- No factorization (ϕ).
- **Community size** ($|c^i|$): the number users registered as members of community c^i.
- **Location popularity** (p^i): the *total* number of pixel allocations on the area defined by l^i, divided by $l^i_{\#pixels}$.
- **Diameter** (d^i): the maximal Manhattan distance between two pixels in l^i.
- **Entropy** (H^i): entropy of the distribution of colors in l^i.

To guide the eye, we illustrate these success measures through four simple toy examples in Figure 3. All four Figures (3a- 3d) are made of 28 pixels (unit cubes), projected on the X-Y plane. However, while having an identical shape, the entropy of Figure 3b is higher than that of Figure 3a, the diameter of Figure 3c is longer than the other logos (14 vs. 8). While the entropy of 3d is equal to that of 3a, the area popularity of Figure 3d is greater than the popularity of the other logos, as some of the 28 pixels were manipulated twice (presumably the yellow tiles were placed by members of one community, then overridden by another community to create the red-black plus-shaped logo).

Formally, taking the naive approach – all Figures are equally successful: $s_\phi(a) = s_\phi(b) = s_\phi(c) = s_\phi(d) = 28$. Taking the diameter into account, we obtain the following order $s_d(c) > s_d(a) = s_d(b) = s_d(d)$, while taking the complexity of the logo into account we get $s_H(b) > s_H(d) > s_H(c) > s_H(a)$, and considering the popularity of (demand for) the location of the logo we get $s_p(d) > s_p(a) = s_p(b) = s_p(c)$.

It is important to note that we find that the diameter positively correlates with the circumference, and therefore it serves as a simple approximation for the number of potential border clashes. Similarly, the definitions of popularity and complexity are only measurable ways that approximate potential dynamics on the canvas.

3 Data

Communities This work takes interest only in the communities that took an active part in the r/place experiment. While a list of 1231 communities was compiled and shared by Israeli, Kremiansky, and Tsur (2022), we find that due to their use of heuristic filters, the list includes some communities that only discussed r/place without participating. We manually verified and filtered all communities in their list, obtaining a subset of 997 well-defined communities that took part in the r/place experiment.

For each community, we obtained a number of meta-features (e.g., age and size $|c^i|$) through Reddit's API[5] as

[5]https://www.reddit.com/dev/api/

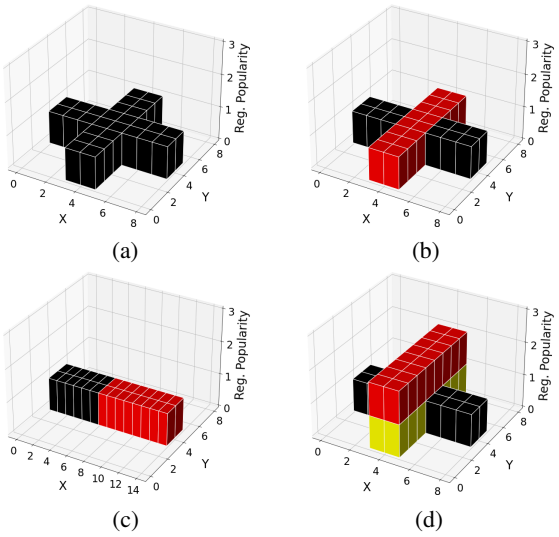


Figure 3: Artwork examples. On the x-y axis, two-dimensional artworks are presented. The z-axis emphasizes how popular each pixel in the artwork was. All four artworks have 28 allocated pixels.

well as all the posts, comments and up/down votes during the 72 hours of the experiment and the three months preceding it. These data are used to generate multifaceted representations of each of the communities, which in turn are used as the input for the prediction models (see Section 4).

We assigned each community five gold labels – one binary and four continuous – each label corresponds to a success measure, as defined in Section 2.

Binary Gold Labels The binary labels were obtained by using the *Place-Atlas*[6] resource for manual annotation. We find that 331 communities (33%) of the participating communities failed to leave a recognizable mark on the canvas.

Continuous Gold Labels The four continuous labels for a community c^i depend on the size of the logo ($l^i_{\#pixels}$), the community size ($|c^i|$), the popularity of the location (p^i), the diameter of the logo (d^i), and the entropy of the logo (H^i). Each logo (l^i) was identified using the Place-Atlas, from which we directly derived $l^i_{\#pixels}$, and computed d^i and H^i. The $|c^i|$ and p^i values were extracted from data shared by Reddit[7] and matched with the Place-Atlas data.

Finally, the label values for community c^i are given by:

$$s_{|c|}(i) \triangleq l^i_{\#pixels} \cdot max\{1 - \rho(|c^i|), \alpha\}$$
$$s_p(i) \triangleq l^i_{\#pixels} \cdot max\{\rho(p^i), \alpha\}$$
$$s_d(i) \triangleq l^i_{\#pixels} \cdot max\{\rho(d^i), \alpha\}$$
$$s_H(i) \triangleq l^i_{\#pixels} \cdot max\{\rho(H^i), \alpha\}$$

[6]Place-Atlas: https://draemm.li/various/place-atlas

[7]`r/place` published data: https://tinyurl.com/4ewtwu8w.

	$s_{\|c\|}$	s_p	s_d	s_H
s_ϕ	0.15 ; 0.41	0.06 ; 0.25	0.88 ; 0.89	0.2 ; 0.24
$s_{\|c\|}$	1.0	0.16 ; 0.18	0.18 ; 0.37	0.07 ; 0.09
s_p		1.0	0.07 ; 0.22	-0.07 ; -0.01
s_d			1.0	0.22 ; 0.17

Table 1: Correlations between the different measures of success. s_ϕ is the success rank based on the number of pixels with no accounting for any relevant factor. Value pairs indicate the Pearson (first value) and the Spearman (second value) correlations.

where ρ denotes the percentile of the value, divided by 100,[8] and $\alpha = 0.1$.[9]

Table 1 presents the correlation between the four measures (and the naive $s_\phi(i) = l^i_{\#pixels}$). We observe a positive, though low, correlation between the measures. The correlation values validate our intuition that the sheer number of pixels should be discounted and that each type of label represents a different nuance of the notion of success.

Temporal Datasets One of the primary goals of this work is to predict the level of success of a community based on a multifaceted representation that reflects different characteristics of the community. We hypothesize that specific traits and characteristics can explain the success or failure of a community to rally for a cause or against an emerging threat. Furthermore, we wish to test whether these characteristics are exhibited in the community's everyday, mundane, activity, or emerge at a time of crisis. In order to test this, we split the data into two distinct datasets: (i) the data generated during three months *before* `r/place` (BP: Before Place), and (ii) the data generated *during* the 72 hours of `r/place` (DP: During Place). General statistics describing BP and DP are presented in Table 2.

4 Experimental Setting

4.1 Community Representation

Communities are multifaceted and can be characterized from different perspectives. To this end, we represent Reddit communities by features of four general types: (i) Textual features, (ii) Meta-features, (iii) Network features, and (iv) Network Embeddings. We calculated each feature described below over BP and over DP independently.

Textual representations We normalize the textual data by lower casing, tokenization, removing punctuation, and converting full URL addresses to their domain name only. We experiment with three types of textual features: (i) Bag-of-words features: We use the *TF-IDF* score (Salton and McGill 1986) per token. Using bigrams/trigrams did not yield any improvement, so we report all BOW results for

[8]That is if $|c^i|$ is in the 13th percentile, $1 - \rho(|c^i|) = 0.87$.

[9]This hyper-parameter is used to prevent the diminishing of the success score. We find 0.1 to be adequate, although other small values could be used to the same effect.

	BP (Before Place, 3 months)				DP (During Place, 3 days)			
	Total	Mean	Median	STD	Total	Mean	Median	STD
Active Users	8.61M	17.26K	2.25K	272.95K	1.1M	2.21K	300	34.91K
Submissions	8.4M	16.84K	1.79K	268.68K	352.4K	706.1	89	11.23K
Comments	121.65M	243.79K	18.31K	3.87M	4.7M	9.43K	823	149.43K
Tokens	3840.8M	7.7M	619.17K	121.93M	14.02M	28.09K	3.08K	444.86K

Table 2: Data statistics. BP spans over the period before `r/place`, while DP spans over the time during `r/place`. The mean, median, and standard deviation are calculated over the subreddits.

the unigram setting only. (ii) LIWC categories: The Linguistic Inquiry and Word Counts (LIWC) dictionary is used to assign words to cognitive and emotional categories (Pennebaker, Francis, and Booth 2001). A vector of LIWC categories represents utterances -– each entry reflects the weight of the corresponding LIWC category in that text. We aggregate all utterances found in the community discussions to represent each community in a single LIWC feature vector, capturing the "vibe" of a community. (iii) Raw text: the community's submissions are concatenated and separated with the [SEP] token to fine-tune a BERT model.

Meta-features Each subreddit can be represented by a series of meta-features. For example, the number of users subscribed to it, the average number of posts per day, the average number of up/down votes per post, the age of the community (days since its creation), etc. We use a total of 25 meta-features per subreddit.

Network features A community can be characterized by the patterns of communication between its members. These interaction patterns could be thought of as a social network in which a direct reply by user u to a post by user v constitutes a directed edge $u \rightarrow v$. These networks provide another perspective on the organizational principles of a community and the dynamics between its members. In total, we use 32 network statistics as features (e.g., #nodes, #edges, avg. and std. of various centrality measures, #triangles).

Community Embeddings While the network features described above were used by Israeli, Kremiansky, and Tsur (2022), we find this approach naive. We, therefore, consider three stronger representations of the social graph:

- SNAP embeddings: Pretrained embeddings of 51.2K Reddit communities, including *all* the communities in our data are shared as part of SNAP (Kumar, Zhang, and Leskovec 2019). The embeddings are generated based on data from Jan 2014 to April 2017. The dimension of the node embeddings is 300. The SNAP embeddings are not computed on the BP and DP datasets.
- Community2vec: We use the community embeddings algorithm (Martin 2017) on BP and DP (independently), obtaining embedding vectors with a dimension of 100.
- Graph2vec: We use the graph social structure of each subreddit to train *an unsupervised* graph-level embeddings using the InfoGraph algorithm (Sun et al. 2019). We use the implementation suggested by Liu et al. (2021) with a learning rate of 0.001, trained over five epochs to yield an embedding vector of size 100 per community.

4.2 Prediction Models

In predicting the binary success label we cast the problem as a binary classification task. For the continuous labels, we formulate the problem as a regression task. We experiment with an array of algorithms ranging from simple logistic/linear regression (Searle and Gruber 2016) to gradient-boosted trees (Friedman 2002), feed-forward neural networks, and transformers (Vaswani et al. 2017). We use a deviance loss function for the GBT and the Random-Forest algorithms, while a binary log-loss is used in fine-tuning the BERT model.

4.3 Experimental Settings

Each subreddit is represented by an array of features as described above (Section 4.1), derived separately from the BP and DP data. We execute all algorithms in an ablation manner, in order to evaluate the contribution of the different feature types. Precision, Recall, F1-score, and AUC scores are reported for the binary setting. We report the RMSE and the adjusted R-square in the regression setting. The classification algorithms optimize the F1-score as precision and recall are equally important, given the task definition. The regression algorithms optimize the squared error. We evaluate all algorithms and settings using stratified 5-fold cross-validation. Neural architectures are restricted to a maximum of ten epochs with early stopping.

5 Results and Analysis

We compare our results with two intuitive univariate linear models: one uses the size of the community as the independent variable, and the other uses the community's age (days since creation, counting back from 31/3/2017).

5.1 Classification Results

The best classification results were obtained by the GBT classifier using all types of features (textual, meta, and network) derived from the DP dataset: an average F1-score of 0.695 and AUC of 0.694, over 5-folds. These results significantly outperform the best baseline (F1-score of 0.53 and AUC of 0.55).

Using the BP data the top-performing model achieves an F1 score of 0.647 and an AUC of 0.645. While these results

Suc.	Dataset	Features	RMSE (↓)	Adj. R^2 (↑)
$s_{\|c\|}$	Ext.-Baseline	$\|c\|$	0.558±0.03	0.045±0.045
		Age	0.576±0.039	-0.015±0.007
	BP	Network	0.535±0.042	0.124±0.063
	DP	All	0.519±0.048	0.291±0.021
s_p	Ext.-Baseline	$\|c\|$	0.736±0.035	0.061±0.028
		Age	0.756±0.023	0.008±0.026
	BP	All	0.628±0.038	0.32±0.046
	DP	All	0.593±0.057	0.421±0.038
s_d	Ext.-Baseline	$\|c\|$	0.884±0.031	-0.002±0.021
		Age	0.869±0.047	0.032±0.068
	BP	BOW	0.744±0.047	0.294±0.071
	DP	All	0.727±0.066	0.348±0.039
s_H	Ext.-Baseline	$\|c\|$	0.732±0.037	0.044±0.036
		Age	0.75±0.028	-0.003±0.016
	BP	All	0.646±0.037	0.255±0.037
	DP	All	0.634±0.051	0.327±0.046

Table 3: Regression results. *Succ.*: the success-definition to be predicted. *External (Ext.) Baseline*: univariate linear model. ↓ (↑) indicates that a lower (higher) value is better. *Adj.*: Adjusted.

are inferior to the results obtained based on the DP data, they are still significantly better than the baseline results.

5.2 Regression Results

Table 3 presents regression results in a 5-fold cross-validation setting. Due to space constraints, we report results only for the univariate baseline models and for the best-performing feature set for each label type and dataset.

Using either BP or DP dataset significantly outperforms both baselines in all four label types. The results obtained over DP consistently outperform those obtained over BP. We further discuss this trend in Section 6.

5.3 Analysis and Social Interpretation

The SHAP explanatory toolkit (Lundberg and Lee 2017) allows quantifying the impact of specific features on the prediction. A high (low) SHAP value indicates the feature's positive (negative) impact on the prediction for a specific instance (community). We use the SHAP aggregate values to derive social insights. SHAP values of six prominent features (two word-tokens, two community meta-features, one network feature, and one LIWC feature) are presented in Figure 4. Higher X-axis values indicate a positive contribution to the model's prediction. The color corresponds to the actual value of the feature. The analysis we provide was done on the DP setting.

Planning, alerting, engaging We observe that high values of the word-feature *plan* are correlated with a positive SHAP value. We manually verified that the frequent use of the word is often used in the context of strategic planning of the community's action. Similarly, the word-feature *under* is used to alert the community members as in "we are [our logo is] under attack". On the other hand, we observe a *negative* correlation between the LIWC category 'incl' (Inclusive[10]) and success in the `r/place` experiment.

Interestingly, we observe that a low *distinct comments to submission ratio* (the number of *users* responding to a submission, not to confuse with *comments to submission ratio*) has a positive effect on the prediction score. Our interpretation suggests that successful campaigns allow focused discussions between community members: comments and discussions are encouraged as long as a discussion thread does not lose focus. Too many *members* commenting on a submission (high feature values) may result in stagnation that harms coordination. Furthermore, high values of the *removed submission ratio* is positively associated with success in the game. Removal of submission in Reddit is allowed by the author himself or by the moderators of the community (i.e., a small set of users that manage the community). These features suggest that (self) moderation is important in large-scale distributed campaigns.

The SHAP values for *num of triangles* provide a complementary perspective: a denser network is related to better performance – reinforcing the sociology scholarship asserting that high clustering facilitates trust and high social capital (Coleman 1988). Combined with our interpretation of the SHAP analysis of the *distinct comments to submission ratio* it suggests that successful communities have their members engaged efficiently in focused discussions, rather than being verbose, creating distractions. This interpretation provides additional nuance to well-established theory, e.g., Backstrom et al. (2006); Cunha et al. (2019).

Differences between definitions of success Previous work by Cunha et al. (2019) identified four different measures of long-term success. In Section 2, we presented and motivated five measures of success in concrete campaigns. In this part, we explore the differences between success measures and the correlated factors by focusing on communities for which performance differs radically across success measures[11]. For example, the logo produced by the *League of Legends* (LoL) gaming community consists of 7114 pixels (see Figure 5). Given that it is one of the largest communities in the corpus ($\sim$ two million members), $s_{|c|}(LoL)$ success level is ranked 644 out of the 666 surviving communities (i.e., that left a recognizable mark on the canvas). However, considering the complexity of the logo, $s_H(LoL)$ ranks the community at 14/666.

In total, we identified 134 communities in which their success rank is radically different according to different success measures (top quartile in one measure and in the bottom quartile in another). We denote this set of communities c^{Δ}.

Table 4 presents the average ablation RMSE for $s_{|c|}$ and s_H on all communities in C^{Δ}. It is evident that different fea-

[10]Examples of words in this category: *with*, *together* and *plus*.

[11]Due to space constraints, we provide analysis only for the s_d (logo diameter) and s_H (complexity) success measures.

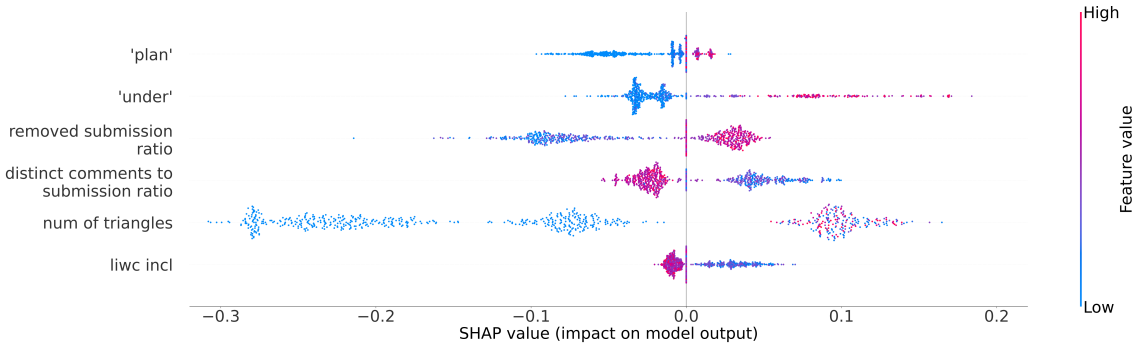


Figure 4: SHAP values for the six prominent features in modeling the s_p label while using the DP dataset.

Figure 5: LeaguOfLegends artwork. The artwork is located at the center of the canvas, left of the France flag.

Feature Type	$s_{\|c\|}$	s_H
Meta	**0.438**	0.51
Network	0.444	0.525
LIWC	0.532	0.549
BOW	0.529	**0.489**
Com2Vec	0.478	**0.5**
SNAP	0.501	0.55
Graph2Vec	**0.427**	0.521
All	0.436	0.504

Table 4: Average RMSE values of $s_{|c|}$ and s_H success prediction for communities in C^{Δ}. Best performing feature types are marked in boldface.

ture types play a more/less significant role, depending on the definition of success. For example, the Meta and Graph2Vec features better predict success when accounting for community size. In contrast, BOW and Com2Vec features are helpful where the focus is on the complexity of the goal.

Using SHAP analysis to recover the role of specific features in predicting $s_{|c|}$, we find that the number of community members, the number of comments posted, and the average upvote score are the most important meta features; the average centrality, the density of the network, and the number of triangles are found to be the best predictors among the network features. This is inline with the findings of Cunha et al. (2016) and Cheng, Danescu-Niculescu-Mizil, and Leskovec (2014) regarding the importance of engagement and positive feedback.

Differences between BP and DP We further analyze the differences between the community signal obtained prior to the experiment and the signal obtained during the experiment. Due space constraints, we present the analysis only over the S_p success label.[12] A comparison between BP and DP is provided in Table 5. [13] We observe that the DP model better predicts success over all types of features. In both cases, the model that combines all types of features performs best.

To find further differences between BP and DP, we analyze the feature importance distribution (by SHAP values) of the models that combine *all* feature types. We find that in both models, the 'BOW' features contribute the most (39.8% and 60.5% in the DP and BP models, respectively). However, the 'Network' features contribute 30.6% of the features' importance in the DP model compared to 14.7% in the BP model. We also observe a stronger contribution of the LIWC features in the DP model (16.2% VS. 12.8%).

We further contrasted the models' predictions in the same community, finding the cases in which the models' predictions differ significantly. Such cases highlight *outlier communities* that over (or under) perform during the `r/place` experiment. Two such outlier communities are r/osugame (see Figure 1, last row) and r/straya (Australia). In both communities, the 'Network' features were significantly more dominant in the DP model compared to the importance of the 'Network' features in the BP model.

Interestingly, national/geographical communities (e.g., r/straya) are over-represented in the set of outlier communities. Out of the top 100 outlier communities, 23% are national/geographical communities, while only 10.6% (106 out of 997) of the participating communities are national/geographical communities. We hypothesize that such an over-

[12]The S_p achieves the best R^2 over all the labels, see Table 3.

[13]Note that the SNAP and Com2Vec features types are not included since they are obtained from external sources that do not provide different representations for BP and DP.

Feature Type	BP	DP	Cor.
Meta	0.26	0.33	0.81
Network	0.27	0.328	0.83
LIWC	0.172	0.269	0.59
BOW	0.301	0.364	0.72
Graph2Vec	0.25	0.29	0.78
All	0.322	0.421	0.8

Table 5: R^2 results (higher is better) for the location popularity success measure (S_p) using BP and DP datasets. Cor. represent the Pearson correlation between the two predictions.

representation is due to the nature of the r/place settings. Flags (and other national symbols) are recognizable by all community users, uncontroversial among most members, and relatively easy to draw. Moreover, it is well established that national identity is one of the stronger totems of personal identity, having individuals unite, fight and protect national symbols (Mudde 2007; Reicher and Hopkins 2000; Smith and Smith 2013; Jaskulowski 2016).

6 Discussion

The limited performance of LLMs State-of-the-art results in many prediction tasks are often achieved by fine-tuning LLMs. We have experimented with a number of LLMs, including distillBERT (Sanh et al. 2019) and the Longformer (Beltagy, Peters, and Cohan 2020) which is more adequate to handle longer sequences of texts. The performance of the LLMs were disappointing[14]. We attribute the modest performance of LLMs on the task and data at hand to two factors: First, the number of instances (communities) is relatively small, which may not be enough for training large models. Second, LLMs capture the topic and semantics of the texts, but these signals are not as important as the community structure and the community dynamics.

Success measures and the community objective We observe a significant difference in the success (gold labels and predicted labels) with respect to the different measures of success. It is interesting to note that in practice, the $s_d(i)$ and $s_H(i)$ values depend on the explicit objective the community members aim to achieve as they decide on the artwork's shape and complexity. On the other hand, the $s_{|c|}(i)$ is not directly controlled by the c^i since the size of each community is mostly fixed prior to r/place (new users cannot join, although registered users can migrate between communities). Finally, $s_p(i)$ is controlled to some degree by c^i: some communities deliberately choose to operate in regions of high demand (e.g., the U.S. flag in the center of the canvas), while other communities (e.g., r/osuGame) operated at the periphery of the canvas just to be repeatedly attacked by TBV. These clashes made the OSU location the most popular area in terms of pixel changes.

Generalizability We use the r/place experiment to model success levels of online communities. The unique setting allows us to consider multiple ways to quantify success. On the one hand, the set of explanatory features and success measures are specific to r/place, while on the other hand, we use general *concepts* that can be used in other experimental settings. For example, the entropy of the artwork and the location on the canvas are r/place-specific, but the complexity of the group's objective and the opposition it faces are general. Similarly, a specific word-token can have a high SHAP value in this context, but using word tokens, community structure, and other features are general enough and could be used for many modeling tasks.

Limitations The nature of the Reddit platform and the r/place experiment attracted specific communities and demographics – a few million users organized in about a thousands of communities make only a small part of the users and the communities[15] on Reddit. Our modeling and analysis only include those who participated.

In this work, we propose different success measures that rely on previous studies as well as the nature of the r/place experiment. However, quantifying success according to the *objective measures* internally defined per community would be a more suitable choice (e.g., blocking another competing community). Recovering goals this specific is extremely challenging and may not even have any clear indication in the data.

7 Related Work

Community dynamics and collaborative action The behaviors, norms, and dynamics of human communities are at the core of the social science research (Lewin 1947b, 1948; Lewin et al. 1947). Naturally, in the last decade, much of the research has been geared towards online communities over social platforms, e.g., (Lazer et al. 2009; Zhang et al. 2017; Mensah, Xiao, and Soundarajan 2020).

Reddit data have been used extensively to study various aspects of the organization, development, evolution, and behavior of online communities. A general overview of the study of Reddit communities is provided by Medvedev, Lambiotte, and Delvenne (2017). While in our work, we study hundreds of communities, other works focus on a *single* community, presenting its uniqueness and norms (Jones et al. 2019; August et al. 2020; Britt et al. 2021).

Evolving community behaviors, the effect of moderation on Reddit communities, and the different factors that cause a community to evolve are studied in a battery of studies (Weninger, Zhu, and Han 2013; Choi et al. 2015; Stoddard 2015; Cunha et al. 2016; Fiesler et al. 2018; Rappaz et al. 2018; Mensah, Xiao, and Soundarajan 2020) to mention just a few. These works address various aspects of community organization as an interest group, the dealings with topics of interest, and the inherent tension between anonymity and

[14]For example, in the binary classification task ("leaving a mark") the BERT model achieved an average F1-score of 0.627 using the BP dataset, compared to 0.647 by the GBT.

[15]There are $\sim 1.7B$ registered users and $\sim 1.2M$ communities, though many are inactive.

identity. Recent works study the structure and other characteristics (e.g., loyalty) of Reddit communities (Zhang et al. 2017; Hamilton et al. 2017; Kumar et al. 2018; Zhou and Jurgens 2020; Massachs et al. 2020). Zhang et al. (2017) suggests a new representation of communities through 'distinctiveness' and 'dynamicity' dimensions. The authors emphasize that these representations reflect different user engagement measures (e.g., retention rate). Kumar et al. (2018) suggests a novel way to model conflicts between online communities. Their approach integrates textual data with communities' meta-features. This methodology is similar in a way to the methodology we use to combine different representations of communities. Datta and Adar (2019) expand this work and study the landscape of conflicts among communities on Reddit.

Success of communities In this work, we model community *success*. A series of studies tackle this topic using multiple definitions of success (Kairam, Wang, and Leskovec 2012; Tan 2018; Cunha et al. 2019). These definitions are contingent upon metrics linked to the temporal evolution of a community's activities. E.g., the number of posts generated, and growth rate, and the members' retention.

In stark contrast, our concept of *success* is fundamentally distinct from these previous approaches. We introduce a quantitative definition of success, which is grounded in the evaluation of a community's performance within the context of an organic, large-scale campaign, necessitating the collective engagement of its members.

Cunha et al. (2019) identify four success measures associated with communities and analyze their relationship. They conclude that success is multi-faceted and can hardly be measured nor predicted by a single measurement. Their work and approach inspire our work. We hypothesize that success in the `r/place` experiment has to be measured using multiple measurements, each capturing a different facet of success. We also assume that a unique predictive model should be fitted per success measure.

The r/place experiment Studies that utilize the `r/place` experiment data are still scarce. Müller and Winters (2018) study the experiment from a perspective of how artworks evolve over time and their correlation with the canvas density. Rappaz et al. (2018) and Armstrong (2018) introduce an analysis of the latent patterns of collaboration between individuals. Conflicts between communities during the `r/place` experiment are studied by (Vachher et al. 2020), which also releases a dataset of conflict regions and the communities involved in each. None of these works tackle the prediction tasks we propose. In addition, they fundamentally differ from our work since they focus on modeling *individual Redditors* while we focus on the *community* level, and they only use the `r/place` pixel allocations data while we combine *multiple types of features* (language, community structure, user dynamics, etc.).

Litherland and Mørch (2021) introduced a framework to analyze specific communities that draw artwork in `r/place`. They study the evolution of visual artifacts and social artifacts during the experiment. However, they use limited data for the analysis (only structural) and focus on a single community (the Mona Lisa painting).

Community engagement in large-scale distributed campaigns was recently studied by (Israeli, Kremiansky, and Tsur 2022). This work is closely related to our work. Both works focus on communities rather than individual users and use a similar computational approach involving the integration of multiple feature types into the prediction model. However, our work differs from Israeli, Kremiansky, and Tsur (2022) in three key aspects: (i) We define a different research question – rather than focusing on the (non)participation of a community in `r/place`, we focus on the *success level* of participating communities, (ii) While Israeli, Kremiansky, and Tsur (2022) use only data that was generated before `r/place`, we also examine the signal produced *while* the experiment. We compared the performance of the models using data generated before and during the experiment, and (iii) We follow Kumar, Zhang, and Leskovec (2019) and use community embeddings, together with 'naive' representations based on a predefined set of features (e.g., centrality).

8 Conclusions and Future Work

We study how community structure, language, internal norms, and other characteristics can be used to predict the success level of a community in large-scale distributed campaigns. Specifically, we predict how well a Reddit community performs in the `r/place` experiment. We argue that success can be defined in a number of ways that are not well correlated. Defining a number of success measures, we experimented with a high number of representations per community (e.g., language and network) – calculated before and during the `r/place` experiment.

We found that the data collected *during* the experiment are more effective than the data collected *before* the experiment, overall and for each feature type separately. We find that certain words, such as *plan* and *under*, are highly correlated with the success level in `r/place`. We also find structural characteristics, such as the number of triangles and the network density, that are positively correlated with success in the game. We find that communities conducting more focused discussions have a better success rate. Finally, relying on a novel comparison between two of our models, we find that the success level of national/geographical communities (e.g., r/straya) is not well predicted by the models, suggesting that these communities have a unique behavior while engaging in the `r/place` experiment.

Future work takes two trajectories: (i) Model the 2022 `r/place` experiment[16] and (ii) Model the behavior and collaboration between *users* in the `r/place` experiment.

Broader perspective: ethics One primary objective of this work is to understand better the factors that contribute to a successful undertaking of a campaign by a community. The insights derived from the predictive models and from the analysis of the results can help communities coordinate in order to promote a cause or mitigate adversarial cam-

[16] A new version of the `r/place` experiment that attracted 16M Redditors but had a very different setting than the 2017 experiment.

paigns. On the other hand, these insights could also be utilized in designing more efficient adversarial campaigns such as election interference, increasing polarization, or promoting fake news or anti-scientific sentiment.

References

Albini, P. 2017. Archive of Reddit's r/place. https://github.com/pietroalbini/reddit-place-2017.

Armstrong, B. 2018. *Coordination in a Peer Production Platform: A study of Reddit's/r/Place experiment*. Master's thesis, University of Waterloo.

August, T.; Card, D.; Hsieh, G.; Smith, N. A.; and Reinecke, K. 2020. Explain like I am a Scientist: The Linguistic Barriers of Entry to r/science. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

Backstrom, L.; Huttenlocher, D.; Kleinberg, J.; and Lan, X. 2006. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 44–54.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Britt, B. C.; Britt, R. K.; Hayes, J. L.; Panek, E. T.; Maddox, J.; and Musaev, A. 2021. Oral healthcare implications of dedicated online communities: A computational content analysis of the r/Dentistry subreddit. *Health communication*, 36(5): 572–584.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2014. How community feedback shapes user behavior. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Choi, D.; Han, J.; Chung, T.; Ahn, Y.-Y.; Chun, B.-G.; and Kwon, T. T. 2015. Characterizing conversation patterns in Reddit: From the perspectives of content properties and user participation behaviors. In *Proceedings of the 2015 ACM on Conference on Online Social Networks*, 233–243. ACM.

Coleman, J. S. 1988. Social capital in the creation of human capital. *American journal of sociology*, 94: S95–S120.

Côté, J. E. 1996. Sociological perspectives on identity formation: The culture–identity link and identity capital. *Journal of adolescence*, 19(5): 417–428.

Cunha, T.; Jurgens, D.; Tan, C.; and Romero, D. 2019. Are all successful communities alike? Characterizing and predicting the success of online communities. In *The World Wide Web Conference*, 318–328.

Cunha, T. O.; Weber, I.; Haddadi, H.; and Pappa, G. L. 2016. The effect of social feedback in a reddit weight loss community. In *Proceedings of the 6th International Conference on Digital Health Conference*, 99–103. ACM.

Datta, S.; and Adar, E. 2019. Extracting inter-community conflicts in reddit. In *Proceedings of the international AAAI conference on Web and Social Media*, volume 13, 146–157.

Fiesler, C.; Jiang, J. A.; McCann, J.; Frye, K.; and Brubaker, J. R. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. In *ICWSM*, 72–81.

Fisher, D. R.; Andrews, K. T.; Caren, N.; Chenoweth, E.; Heaney, M. T.; Leung, T.; Perkins, L. N.; and Pressman, J. 2019. The science of contemporary street protest: New efforts in the United States. *Science advances*, 5(10): eaaw5461.

Friedman, J. H. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4): 367–378.

Granovetter, M. S. 1973. The strength of weak ties. In *Social networks*, 347–367. Elsevier.

Hamilton, W. L.; Zhang, J.; Danescu-Niculescu-Mizil, C.; Jurafsky, D.; and Leskovec, J. 2017. Loyalty in online communities. In *Eleventh International AAAI Conference on Web and Social Media*.

Hässler, T.; Ullrich, J.; Bernardino, M.; Shnabel, N.; Valdenegro, D.; Van Laar, C.; Sebben, S.; Visintin, E.; Tropp, L.; González, R.; et al. 2020. A large-scale test of the link between intergroup contact and support for social change. *Nature Human Behaviour*.

Israeli, A.; Kremiansky, A.; and Tsur, O. 2022. This Must Be the Place: Predicting Engagement of Online Communities in a Large-scale Distributed Campaign. In *Proceedings of the ACM Web Conference 2022*, 1673–1684.

Jackson, S. J.; and Foucault Welles, B. 2016. # Ferguson is everywhere: initiators in emerging counterpublic networks. *Information, Communication & Society*, 19(3): 397–418.

Jaskulowski, K. 2016. The magic of the national flag. *Ethnic and Racial Studies*, 39(4): 557–573.

Jones, R.; Colusso, L.; Reinecke, K.; and Hsieh, G. 2019. r/science: Challenges and opportunities in online science communication. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–14.

Kairam, S. R.; Wang, D. J.; and Leskovec, J. 2012. The life and death of online groups: Predicting group growth and longevity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 673–682.

Kumar, S.; Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference*.

Kumar, S.; Zhang, X.; and Leskovec, J. 2019. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1269–1278. ACM.

Lazer, D.; Pentland, A. S.; Adamic, L.; Aral, S.; Barabasi, A. L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; et al. 2009. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915): 721.

Lewin, K. 1947a. Frontiers in Group Dynamics: Concept, Method and Reality in Social Science; Social Equilibria and Social Change. *Human Relations*, 1(1): 5–41.

Lewin, K. 1947b. Frontiers in group dynamics: II. Channels of group life; social planning and action research. *Human relations*, 1(2): 143–153.

Lewin, K. 1948. *Resolving social conflicts; selected papers on group dynamics*. Harper.

Lewin, K.; et al. 1947. Group decision and social change. *Readings in social psychology*, 3(1): 197–211.

Litherland, K. T.; and Mørch, A. I. 2021. Instruction vs. emergence on r/place: Understanding the growth and control of evolving artifacts in mass collaboration. *Computers in Human Behavior*, 122: 106845.

Liu, M.; Luo, Y.; Wang, L.; Xie, Y.; Yuan, H.; Gui, S.; Yu, H.; Xu, Z.; Zhang, J.; Liu, Y.; et al. 2021. Dig: A turnkey library for diving into graph deep learning research. *arXiv preprint arXiv:2103.12608*.

Lucchini, L.; Aiello, L. M.; Alessandretti, L.; Morales, G. D. F.; Starnini, M.; and Baronchelli, A. 2021. From Reddit to Wall Street: The role of committed minorities in financial collective action. *arXiv preprint arXiv:2107.07361*.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 4765–4774.

Mancini, A.; Desiderio, A.; Di Clemente, R.; and Cimini, G. 2022. Self-induced consensus of Reddit users to characterise the GameStop short squeeze. *Scientific reports*, 12(1): 1–11.

Martin, T. 2017. community2vec: Vector representations of online communities encode semantic relationships. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, 27–31.

Massachs, J.; Monti, C.; Morales, G. D. F.; and Bonchi, F. 2020. Roots of trumpism: Homophily and social feedback in donald trump support on reddit. In *12th ACM Conference on Web Science*, 49–58.

McMillan, D. W.; and Chavis, D. M. 1986. Sense of community: A definition and theory. *Journal of community psychology*, 14(1): 6–23.

Medvedev, A. N.; Lambiotte, R.; and Delvenne, J.-C. 2017. The anatomy of Reddit: An overview of academic research. In *Dynamics on and of Complex Networks*, 183–204. Springer.

Melucci, A. 1996. *Challenging codes: Collective action in the information age*. Cambridge University Press.

Mensah, H.; Xiao, L.; and Soundarajan, S. 2020. Characterizing the Evolution of Communities on Reddit. In *International Conference on Social Media and Society*, 58–64.

Mudde, C. 2007. *Populist radical right parties in Europe*. Cambridge university press.

Müller, T. F.; and Winters, J. 2018. Compression in cultural evolution: Homogeneity and structure in the emergence and evolution of a large-scale online collaborative art project. *PloS one*, 13(9): e0202019.

Olson, M. 2009. *The Logic of Collective Action: Public Goods and the Theory of Groups, Second printing with new preface and appendix*, volume 124. Harvard University Press.

Ostrom, E. 2000. Collective action and the evolution of social norms. *Journal of economic perspectives*, 14(3).

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001): 2001.

Peters, G.; Portman, R.; Klobuchar, A.; and Blunt, R. 2021. Examining The U.S. Capitol Attack: a review of the security planning and response failures.

Rappaz, J.; Catasta, M.; West, R.; and Aberer, K. 2018. Latent structure in collaboration: the case of Reddit R/place. In *Twelfth International AAAI Conference on Web and Social Media*.

Reicher, S.; and Hopkins, N. 2000. *Self and nation*. Sage.

Salton, G.; and McGill, M. J. 1986. *Introduction to modern information retrieval*. McGraw-Hill, Inc.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Searle, S. R.; and Gruber, M. H. 2016. *Linear models*. John Wiley & Sons.

Smith, A. D.; and Smith, A. 2013. *Nationalism and modernism*. Routledge.

Stoddard, G. 2015. Popularity Dynamics and Intrinsic Quality in Reddit and Hacker News. In *ICWSM*, 416–425.

Sun, F.-Y.; Hoffmann, J.; Verma, V.; and Tang, J. 2019. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000*.

Tan, C. 2018. Tracing community genealogy: how new communities emerge from the old. In *Twelfth International AAAI Conference on Web and Social Media*.

Todorov, G. 2022. 70+ Important Reddit Statistics 2022. *thrivemyway (Date accessed: 17/10/2022)*.

Vachher, P.; Levonian, Z.; Cheng, H.-F.; and Yarosh, S. 2020. Understanding Community-Level Conflicts Through Reddit r/place. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, 401–405.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Weninger, T.; Zhu, X. A.; and Han, J. 2013. An exploration of discussion threads in social news sites: A case study of the reddit community. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, 579–583. IEEE.

Zachary, W. W. 1977. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4): 452–473.

Zhang, J.; Hamilton, W. L.; Danescu-Niculescu-Mizil, C.; Jurafsky, D.; and Leskovec, J. 2017. Community identity and user engagement in a multi-community landscape. In *11th Inter. AAAI Conference on Web and Social Media*.

Zhou, N.; and Jurgens, D. 2020. Condolences and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 609–626.

8.1 Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Not Relevant
   (g) Did you discuss any potential misuse of your work? Yes
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? No
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes
   (b) Have you provided justifications for all theoretical results? Partly
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? No
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? No
   (e) Did you address potential biases or limitations in your theoretical framework? To some degree
   (f) Have you related your theoretical results to the existing literature in social science? Yes
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? Not Relevant
   (b) Did you include complete proofs of all theoretical results? Not Relevant
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? All possible data and code
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Did not report but applied cross-validation techniques to address this.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No. Not too relevant for this study
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? No. Not too relevant for this study
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes
   (c) Did you include any new assets in the supplemental material or as a URL? Yes
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No. It does not at all
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Not relevant
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? No. A datasheet is not relevant to these type of data. See data description in the paper (Section 3) and the shared GitHub repo.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? Not Relevant
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Not Relevant
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Not Relevant
   (d) Did you discuss how data is stored, shared, and deidentified? Not Relevant