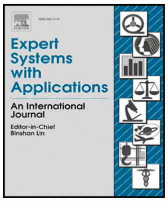
Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Unsupervised discovery of non-trivial similarities between online communities

Abraham Israeli [*], Shani Cohen, Oren Tsur

Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel

ARTICLE INFO

Keywords:
Machine learning
Online communities
Natural language processing
Word embeddings
Social network analysis
Computational social science

ABSTRACT

Language is used differently across communities. The differences may be manifested in vocabulary, style, and semantics. These differences enable the exploration of nuanced similarities and differences between communities. In this work, we introduce C3 — a novel unsupervised approach for community comparison. C3 creates contextual pairwise representations by aligning communities and tuning word embeddings according to both the lexical context and the social context reflected by the community's structure and the community engagement patterns. Specifically, C3 takes into account the semantic relations between pairs of words, reflected by the embeddings model of each community, and leverages the social context and users' role in their community to calculate a similarity measure between community pairs. C3 is evaluated over a dataset of 1565 active Reddit communities, comparing results against three competitive models. We show through an array of experiments and validations that C3 recovers nuanced and not-trivial similarities between communities that are not captured by any of the competitive models. We complement the quantitative results with a qualitative analysis, discussing recovered non-trivial similarities between community pairs such as: *opiates* and *adhd*, *babyBumps* and *depression*, *wallStreetBets* and *sandersForPresident*, all of which are recovered by C3 but not by any of the other models. This qualitative analysis demonstrates the exploratory power of our model.

1. Introduction

The sense of community plays a fundamental role in our life, shapes our identity, increases satisfaction, and reduces stress (McMillan & Chavis, 1986). Communities are organized via shared values, norms, geographic location, identity or interests (Cheong et al., 2009; Eisenstein et al., 2010).

A comparative study of community structure, norms, language, and dynamics is of major importance in an array of research fields ranging from sociology and anthropology (Hofstede, 2001; Kim & McKenry, 1998; Van den Berg & Wilderom, 2004), to psychology (Hanel et al., 2019; Hornsey & Hogg, 2000), education (Tinto & Love, 1995; Zhang & Sun-Keung Pang, 2016), social learning (Lin & Vassar, 2009), and political science (Soon & Kluver, 2007; Stier et al., 2017), to mention just a few.

The similarities between online communities in terms of language, topics of interest, members, or activity patterns are addressed by Hamilton et al. (2017), Hessel et al. (2016), Kumar et al. (2018), Martin (2017) and Waller and Anderson (2019). These works rely on either textual representations or on users and network representations when measuring the similarity between online communities. However, none of these works combine both representations into a unified algorithmic framework. Moreover, these methods are often designed to study similarities on a specific predefined axis, e.g., distinctiveness (Waller & Anderson, 2019), loyalty (Hamilton et al., 2017), or aggressiveness (Kumar et al., 2018; Zhang et al., 2017), rather than providing a generic and robust exploratory tool that recovers nuanced and non-trivial similarities – a useful capability for both social platform moderators and social science researchers.

To illustrate the intricacies of the challenge of discovering nuanced similarities, consider the following seven Reddit communities: *weddingPlanning*, *babyBumps* (pregnancy related advice), *prolife* (anti-abortion agenda), *wallStreetBets*, *teslaMotors*, *sandersForPresident* and *depression*. While *babyBumps* and *prolife* are both related to pregnancy (as evident by a simple text-based comparison) – they are very different in terms of purpose, tone, and dynamics; *weddingPlanning* and *babyBumps* have

[*] Corresponding author.
E-mail addresses: isabrah@post.bgu.ac.il (A. Israeli), shanisa@post.bgu.ac.il (S. Cohen), orentsur@bgu.ac.il (O. Tsur).
URL: https://www.naslab.ise.bgu.ac.il/orentsur (O. Tsur).

[1] A short squeeze campaign promoted in the *wallStreetBets* subreddit incurred losses of billions to some hedge funds in just a few days during January and February of 2021.

https://doi.org/10.1016/j.eswa.2022.117900
Received 23 February 2022; Received in revised form 15 June 2022; Accepted 15 June 2022
Available online 22 June 2022

a significant overlap of users, while they differ in topic and tone. The *babyBumps* and *depression* communities do not have a significant overlap of users nor high topical similarity, but they are intuitively related as both are dedicated to advise and support, thus they may share engagement patterns, or some topical vocabulary (e.g., due to post-partum depression). On the other hand, communities like *teslaMotors*, *sandersForPresident*, and *wallStreetBets* would not appear intuitively similar. However, the *GameStop* short squeeze[1] and the following events suggest a possible connection between these communities that may have been discovered before the events, and provides some explanations in retrospect (Long et al., 2021). We further discuss the relations between these communities in Section 5.

In this paper, we present C3 (Contextual Community Comparison) – a novel method for community comparison, leveraging both language and community structure. Our method allows an unsupervised distance learning between pairs of candidate communities. As opposed to other methods in which a model is trained to classify communities on a specific axis, C3 provides a generic framework that recovers nuanced similarities between communities in an unsupervised manner. At the core of our method, we quantify the distances between *contexts*. Contexts are defined on multiple levels: the relation between the embeddings of *word pairs* within a specific community, reweighed by the social context computed based on the community structure and engagement patterns. These contextual representations are aligned across communities, allowing us to obtain a contextualized similarity measure. We demonstrate how these contextualized similarities balance topic, style, and community dynamics, thus providing a powerful tool for exploratory analysis at scale.

We evaluate our model via different metrics, against three competitive models based on language and community embeddings. We consider thousands of Reddit communities, demonstrating the effectiveness of our method. Among other things, we demonstrate the uniqueness of our model in discovering non-trivial similarities, undetected by other models.

We complement our evaluation with a qualitative analysis, discussing recovered non-trivial similarities between community pairs such as *opiates:adhd*, *babyBumps:depression*, *gay:socialAnxiety*, and *wallStreetBets:sandersForPresident*, all of which are not recovered by other models.

The remainder of the paper is organized as follows: in Section 2 we provide a brief review of the relevant literature. In Section 3 we describe the data collection process and annotation protocol. In Section 4 we introduce the C3 method in detail. Quantitative results followed by the qualitative analysis are provided in Section 5. Section 6 is dedicated to discuss some of the unique properties of C3.

2. Related work

In the last two decades, social networks have become the dominant written-communication platforms,[2] and so the research about them is consistently rising. Works about social media cover a very large scope — ranging from the general usage of social platforms (Fuchs, 2021; Van & Johannes, 2012) to more specific studies, e.g., fake news detection on social platforms (Sahoo & Gupta, 2021; Shu et al., 2017), mental behavior analysis and modeling on social platforms (Abd Rahman et al., 2018; Bouarara, 2021), and bibliometric analysis on Twitter specifically (Noor et al., 2020; Zhang & Wang, 2018).

In the remaining of the current section, we focus on *communities* on social platforms. We cover works around: (i) Language use in online communities; (ii) Representation and comparison methods of online communities; and (iii) Communities as textual corpora – a specific way to handle online communities' content.

Communities and language-use. The unique dialect and linguistic patterns used by different communities and social groups were recently studied by Blodgett et al. (2016), Del Tredici and Fernández (2017), Eisenstein (2013), Jurgens et al. (2017), Tran and Ostendorf (2016) and Lucy and Bamman (2021). Others address the ways the community's language changes over time and between users (Danescu-Niculescu-Mizil et al., 2013; Huffaker et al., 2006; Nguyen & Rose, 2011; Sankoff & Blondeau, 2007), and see Nguyen et al. (2016) for a comprehensive survey. Recent studies, closer to our domain, model various characteristics that are inherent to the community's organization and its activity. Linguistic and structural features are used by Hamilton et al. (2017) to predict members' loyalty. The aggressiveness of a community and its tendency to engage in conflicts with other communities is modeled by Kumar et al. (2018), Zhang et al. (2017) and Datta and Adar (2019).

Community representations and comparison. Formal (vectorial) representations of communities encode a range of signals. Community embeddings based on the intersection between members of different communities are learned by Martin (2017) and Waller and Anderson (2019), while the interaction network within a community is used by Hamilton et al. (2017). Language-based representations are common (Del Tredici & Fernández, 2017; Tran & Ostendorf, 2016), among others. These representations are often designed to capture, and allow prediction of, specific traits and actions, whether it is the distinctiveness of a community (Zhang et al., 2017) or the aggressiveness of a community in conflict with other communities (Kumar et al., 2018). Recommender systems often use similarities between communities in order to prompt users with communities of interest (Janchevski & Gievska, 2019; Olson & Neal, 2015; Spertus et al., 2005). While all these representation methods could be used to measure pairwise similarity, they may produce a biased comparison, capturing similarity on a specific axis (e.g., topical, loyalty, toxicity, or dialect) foregoing other nuances, as shown in Section 5.

Focusing on a single facet, whether it is the content, the structure, or the dynamics, may limit the perspective through which a meaningful comparison could be made. For example, using the user embeddings as suggested by Martin (2017) or by Waller and Anderson (2019) would not allow a comparison between communities from different platforms, since user names cannot be matched, nor a comparison between communities in platforms that maintain the anonymity of community's members. These comparisons are supported by C3.

Communities as textual corpora. A naive, yet straightforward, way to represent communities is to treat the texts produced by each community as a unique textual corpus, each represented in a unique embedding model. Naturally, comparison between, and alignment of, embedding models are widely used in machine translation (Artetxe et al., 2016; Lu et al., 2015; Smith et al., 2017). The relations between corpora across time are studied by Caliskan et al. (2017) as well as by Lewis and Lupyan (2020) and the different biases inherent to different corpora are modeled by Garg et al. (2018). These works measure the distances between words within and across corpora. Gonen et al. (2020) compare embedding models to find differences in the usage of specific words. The comparison is done through an examination of the words' closest neighbors over different embedding models. Most recently, BERT (Devlin et al., 2018) has been used by Lucy and Bamman (2021) to characterize English variation across communities, considering the different senses assigned to words across communities. C3 is inspired by these works. However, we derive features from the community and the discussion structure in order to learn different contextual weights that provide the context for the text-based embeddings.

Contextual representations. C3 relies on multiple types of contexts in computing the similarity between communities. One must not confuse our use of the term 'contextual representations' with text-based contextual embeddings, e.g., ELMo (Peters et al., 2018), the BERT Transformer (Devlin et al., 2018), and other variants of the Transformer architecture.

[2] Facebook reported on 2.9 Billion monthly active users (retrieved 05/31/2022), see: https://tinyurl.com/2p8r4wd6.

Table 1
Central statistical measures. C denotes the final set of sampled communities used in this paper, and C^+ denotes all active communities during a six-month period (10/2016–03/2017). Mean values were averaged over all community submissions. *Community Age* denotes the number of days since the community was established.

	C (1565)				C^+ (258.8 K)			
	Total	Mean	Median	STD	Total	Mean	Median	STD
Subscribers	561 M	377.1 K	39.6 K	1.92 M	1.7 B	6.6 K	30.0 K	234.6 K
Active users	22.1 M	14.1 K	4.4 K	38.5 K	54.7 M	211.4	2.0	5.9 K
Community age	–	2.17 K	2.34 K	0.88 K	–	0.98 K	0.87 K	0.8 K
Submissions	20.7 M	13.2 K	3.6 K	56.5 K	53.3 M	209.1	3.0	6.1 K
Comments per submission	–	12.7	9.4	22.7	–	1.4	0.25	23.9

3. Data

Reddit. Reddit, originally a link-sharing platform, evolved into an active system of dedicated forums. Forums' URLs are marked with a 'r/' followed by the forum's name, and are therefore called 'subreddits'. Subreddits are usually topical, although topics range widely in specificity, from the very broad *r/music* or *r/politics* subreddits to the more specific *r/LilNasX* or *r/sandersForPresident*. Reddit users (called *redditors*) can start a new discussion thread, add a comment to a thread, up/down-vote posts, etc. Following Hamilton et al. (2017), Hessel et al. (2016), Waller and Anderson (2019), Zhang et al. (2017) and Lucy and Bamman (2021) among others (see Section 2), we view each subreddit as a community, having its own internal language style, interaction dynamics, and norms.

Community corpus. The initial dataset consisted of all content (text and meta-data) posted on Reddit during a six-month period (10.1.2016–3.31.2017) – a total of 258.8K active communities. The dataset is available as part of the Pushshift repository (Baumgartner et al., 2020). We denote this set of active communities C^+. We sampled 2580 communities ($\sim$1%) of C^+ for further analysis. From the subset of 2580 communities, we filtered out communities that *did not* meet any of the following three criteria:

1. The language is predominantly English.
2. The vocabulary size is bigger than 5K.
3. At least a third of each community's lexical items occur over five times in that community (accounting for extremely long-tailed word distribution).

This filtering process resulted in C – a set of 1565 communities, revolving around a variety of topics in different granularities (e.g., sports, politics, food, health, pregnancy, depression, herpes, etc.). Further statistics and a comparison between C and C^+ are provided in Table 1.

Community categorization. The subreddit *r/listOfSubreddits* maintains a directory of the subreddits, their category, and their subcategory. For example, *r/bostonCeltics* is filed under Sports (main) and Basketball (subcategory). We validated the category assignments of the subreddits listed in the directory and added labels to 1073 subreddits that were not listed. We make this annotated corpus available in the project's repository.[3] The categories and further statistics regarding the categorization are provided in Table 2.

4. Computing C3

This section describes C3 in detail. For convenience, all the notations used throughout the rest of the paper are summarized in Table 3. There are two main stages in deriving the C3 representations. These representations are then used to compare all community pairs, facilitating the exploratory analysis described in Section 5.

Table 2
Annotated Data Statistics. Subcategory values are the unique number of subcategories defined per category. Each community in C was annotated with a single category and a single subcategory. For example, *r/hawks* is the community of the Chicago Blackhawks hockey team. Hence, it was annotated with 'Sports' as the category and 'Hockey' as the subcategory.

Category	Sub-categories	Total communities	Example
Sports	6	169	r/hawks
Advice/Sharing	8	144	r/lifeprotips
Education	3	83	r/math
Culture/Art	3	46	r/stephenking
Technology	7	139	r/howtohack
Entertainment	6	176	r/pokemon
Video Games	10	295	r/nier
Music	4	84	r/sadboys
Lifestyle	5	98	r/islam
Geographic	3	152	r/serbia
News/Politics	2	60	r/usanews
Hobby/Profession	9	115	r/flyfishing

Table 3
Notations of main variables used in this paper.

Notation	Explanation	Notation	Explanation
c	A specific community (e.g., *r/bostonCeltics*)	C	Full communities corpus, $\|C\| = 1565$
w	A token (usually a word)	W	Set of tokens
V	Full vocabulary	V_c	Full vocabulary of community c
E_c	Embedding matrix of community c	$e_{w,c}$	Embedding vector of token w in the embedding matrix of community c

4.1. Stage 1: Community-specific language modeling

This initial stage is straight forward – a distinct word embedding model E_c is learned for *each* community $c \in C$. For simplicity, we use CBOW (Mikolov, Le, & Sutskever, 2013; Mikolov, Sutskever, et al., 2013), however other embedding algorithms can be used. These distinct embeddings, tuned for the textual context of a community at large, are used in the next stage, in which word contexts, community structure, and communication patterns serve to establish a unique relation between community *pairs* – the core of the C3 framework.

4.2. Stage 2: Contextual-pairwise alignment and comparison

Given two communities $\{c_i, c_j\} \in C$ and their embeddings $\{E_{c_i}, E_{c_j}\}$, as learnt in Stage 1, we create a weighted-contextualized alignment, leveraging both language and community structure. These alignments are created in six steps (A–F), as detailed below and illustrated in Fig. 1.

Step A: Shared vocabulary. We recover the set of tokens (words) used in both communities c_1 and c_2:

$$W_{c_1 \cap c_2} = \{w_i | w_i \in c_1 \wedge w_i \in c_2\}$$

[3] https://github.com/NasLabBgu/C3-contextual-community-comparison.

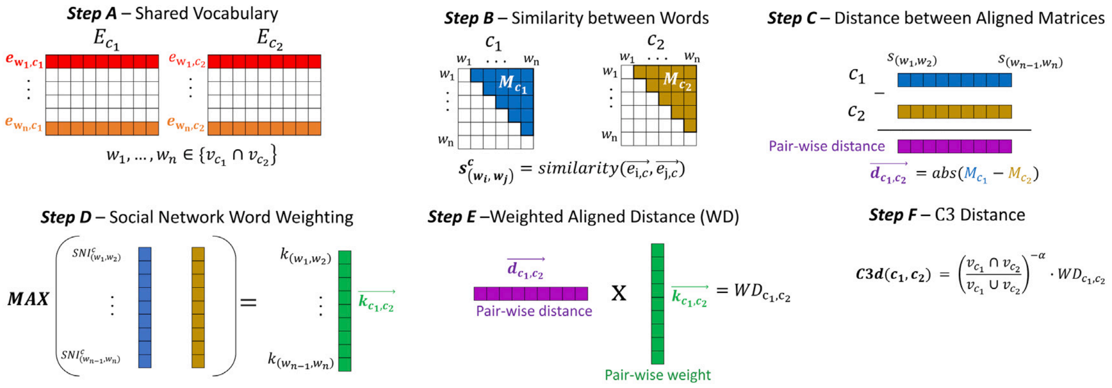

Step A – Shared Vocabulary
E_{c_1}
E_{c_2}
e_{w_1,c_1}
e_{w_1,c_2}
e_{w_n,c_1}
e_{w_n,c_2}
$w_1, \dots, w_n \in \{v_{c_1} \cap v_{c_2}\}$
Step B – Similarity between Words
c_1
c_2
w_1 . . . w_n
M_{c_1}
M_{c_2}
$s^c_{(w_i,w_j)} = similarity(\overrightarrow{e_{i,c}}, \overrightarrow{e_{j,c}})$
Step C – Distance between Aligned Matrices
$s_{(w_1,w_2)}$
$s_{(w_{n-1},w_n)}$
Pair-wise distance
$\overrightarrow{d_{c_1,c_2}} = abs(M_{c_1} - M_{c_2})$
Step D – Social Network Word Weighting
MAX
$SNI^c_{(w_1,w_2)}$
$SNI^c_{(w_{n-1},w_n)}$
$k_{(w_1,w_2)}$
$k_{(w_{n-1},w_n)}$
$\overrightarrow{k_{c_1,c_2}}$
Step E –Weighted Aligned Distance (WD)
$\overrightarrow{d_{c_1,c_2}}$
Pair-wise distance
X
$\overrightarrow{k_{c_1,c_2}} = WD_{c_1,c_2}$
Pair-wise weight
Step F – C3 Distance
$C3d(c_1,c_2) = \left(\frac{v_{c_1} \cap v_{c_2}}{v_{c_1} \cup v_{c_2}}\right)^{-\alpha} \cdot WD_{c_1,c_2}$

Fig. 1. The six steps in Stage 2 of C3. The figure illustrates Steps A to F, over a single pair of communities, c_1 and c_2. This stage of C3 gets as input word embedding models of both communities (E_{c_1} and E_{c_2}) in the format of embedding matrices and returns a distance measure between the two communities. Step D utilizes the social network data in order to assign weights to each pair of words.

Step B: Similarity between words. Per each community c_1 and c_2 we *independently* compute the cosine similarity between the embedding vectors for *each pair* of words $\{w_i, w_j\} \in W_{c_1 \cap c_2}$ as follows:

$$s^c_{(w_i,w_j)} = \frac{\overrightarrow{\mathrm{e_{i,c}}} \cdot \overrightarrow{\mathrm{e_{j,c}}}}{\|\overrightarrow{\mathrm{e_{i,c}}}\| \|\overrightarrow{\mathrm{e_{j,c}}}\|}$$

The $s^c_{(w_i,w_j)}$ measure is computed for all (w_i, w_j) pairs of words in each of the communities, resulting in two dense matrices M_{c_1} and M_{c_2}. $M^{i,j}_c$ holds the similarity between w_i and w_j in community c. Since M_{c_1} and M_{c_2} are defined for the same set of words, these matrices provide contextual alignment of the semantic relations between word pairs that are used in the respective communities.

Step C: Distance between aligned matrices. Given the M_{c_1} and M_{c_2} matrices from step B, we compute the distance vector between each corresponding element i, j above the diagonal in the matrices:

$$\overrightarrow{\mathrm{d_{c_1,c_2}}} = |M_{c_1} - M_{c_2}|$$

For convenience, these elements above the diagonal are illustrated in Fig. 1 as a one dimensional vectors, in which the lines above the diagonal are concatenated. The resulting distance vector $\vec{\mathrm{d}}$ represents the *unweighted* distance between community c_1 and community c_2.

Step D: Social network word weighting. We view each community as a social network with users as nodes and direct correspondence as weighted directed edges. We now assign each word $w_i \in W_{c_1 \cap c_2}$ with a social network importance' score ($SNI^c_{w_i}$) – capturing the overall weight of the word in community c. The SNI score reflects three factors: (i) The frequency of the word's usage in the community, (ii) The inverted frequency of the word in the dataset (each community as a corpus), and (iii) The social status of the community members using it. Thus, we define the SNI score as:

$$SNI^c_{w_i} = \sum_{u \in U_c} freq(w_i, u, c) \cdot Imp^c_u \cdot IDF^C_{w_i}$$

where Imp^c_u denotes the status (importance) of a user u in community c, approximated by average number of up-votes u achieves in the community, thus: $Imp^c_u = \frac{1}{|P_{u,c}|} \sum_{p \in P_{u,c}} UpVotes(p)$, where $P_{u,c}$ denotes the set of posts user u published in community c.

Note that a user posting in more than one community is assigned a different importance score in each of the communities, reflecting her status differences in each of the communities. Consequently, a specific word w_i is expected to have a different SNI score in different communities. The $\overrightarrow{\mathrm{SNI_c}}$ holds the SNI values for the words used in community c.

Remember that while considering two communities c_1 and c_2, the vectors $\overrightarrow{\mathrm{SNI_{c_1}}}$ and $\overrightarrow{\mathrm{SNI_{c_2}}}$ are calculated only for words appearing on both c_1 and c_2. We can thus create a single vector $\overrightarrow{\mathrm{k_{c_1,c_2}}}$ by taking the pairwise maximum values of the respective entries in the $\overrightarrow{\mathrm{SNI}}$ vectors, see illustration in Fig. 1(D). Our choice of SNI weighting, compared to other weighting functions was validated using the self-similarity sanity check described in Section 5.3. Detailed comparison is provided in Section 4.3.

Finally, we normalize $\overrightarrow{\mathrm{k_{c_1,c_2}}}$ using the following normalization function:

$$k^*_{(w_i,w_j)} = \frac{k_{(w_i,w_j)}}{\sum k_{(w_i,w_j)}}$$

This vector will be used in the next step to re-weigh the $\overrightarrow{\mathrm{d_{c_1,c_2}}}$ vector produced in the previous step (C).

Step E: Weighted aligned distance (WD). The output of Steps C and D are used in order to compute a scalar — the weighted distance factor between communities c_1 and c_2:

$$WD_{c_1,c_2} = \overrightarrow{\mathrm{d_{c_1,c_2}}} \cdot \overrightarrow{\mathrm{k_{c_1,c_2}}}$$

Step F: C3 distance. Finally, the C3 distance measure between communities c_1 and c_2 is defined as:

$$\mathrm{C3}_d(c_1, c_2) = \phi \cdot WD_{c_1,c_2}$$

where ϕ is a prior reflecting the 'vocabulary agreement', defined as:

$$\phi = \left(\frac{V_{c_1} \cap V_{c_2}}{V_{c_1} \cup V_{c_2}}\right)^{-\alpha}$$

We introduce ϕ, in order to push apart communities that share only a limited vocabulary (see Step A), while α controls the impact of the topical similarity, reflected by a shared vocabulary. For simplicity, we set $\alpha = 1$.

4.3. C3 parameters

Words weighting evaluation. Step D in Stage 2 of C3, consist of words weighting (see Section 4.2). We examined three alternatives: (i) no weighting, (ii) TF–IDF based weighting, and (iii) social network importance (SNI). The TF–IDF based weighting uses the original TF–IDF logic (Salton & Buckley, 1988), with a minor adaptation for the context of communities (rather than documents/posts). For calculating the TF–IDF, we view the full-textual content of a community as a long document, thus the TF–IDF weight can be easily calculated per token in each community. For the social network importance, we experiment with a number of alternatives (e.g., users' centrality, users' seniority, and users' activity) — all of which take into account the importance of a user within the social network recovered from the interactions between community members. The results we report are based on the average number of up-votes per user — the best performing SNI measure in the evaluation process that we describe next.

Table 4
Evaluation of three weighting options. Communities are compared to themselves, by splitting each into two parts and calculating a normalized distance between the two parts (lower is better). The numbers presented are the average distances over all communities.

Weighting alternative	Average self-distance
No weighting	0.101
TF–IDF weighting	0.084
Social network importance (SNI)	**0.073****

Statistical significance of the best result is indicated by stars (** indicates a P-value $< 10^{-4}$), based on U-test hypothesis testing.

In order to evaluate the three weighting methods listed above, we used the self-similarity evaluation process that is described in Section 5.3. The top-1 error rate is extremely low over all weighting alternatives (ranges between zero and two). Hence, we evaluated the weighting alternatives using the normalized distance of the analyzed community and its supplemental half (lower is better). The average distance measure over a development dataset (10% of the 1565 communities corpus) is presented in Table 4. The SNI outperforms other alternatives (p-value $< 10^{-4}$), based on U-test hypothesis testing (Mann & Whitney, 1947).

Maximum vs. Average in $\vec{\mathrm{k}}$. As described in Section 4.2, we take the maximum SNI values. We experiment with different aggregation functions (e.g., average, minimum) but decided to use the maximum, allowing the words in the more dominant community to control the distance measure.

5. Experimental setting and results

The general similarity between communities (in contrast to the similarity of specific predefined aspects such as loyalty, toxicity, topic, size, etc.) may appear subjective. Therefore, in order to demonstrate the validity and contribution of C3, we report and analyze results in an array of settings, compared to three other models. The next Section 5.1 describes the competing models in detail.

In Section 5.2 we report on the correlation between the different models, establishing that all models provide decent results. We then measure self-similarity as a sanity check, demonstrating that C3 performs better than other methods – a strong indicator for the validity and the power of our model (see Section 5.3). The premise of this work is that C3 captures nuanced and non-trivial similarities, overlooked by other methods. We demonstrate this unique ability in Section 5.4 by evaluating the performance in a carefully designed annotation task. Finally, in Section 5.6 we provide a qualitative analysis of some of the results, yet again, demonstrating the benefits of using C3 for exploratory analysis.

5.1. Baseline models

We compare C3 to three other models: (i) Modified TF–IDF, (ii) Doc2Vec — a text-based approach proven useful in the classification of long texts part of large corpora, and (iii) Com2Vec — a user-embeddings approach used for a direct comparison between communities.

Modified TF-IDF. The model treats the texts produced in a community as a single document in a corpus. The various communities compose the corpus. Documents are represented by the TF–IDF values over a vocabulary and a similarity score is assigned to pairs of documents (communities). This simple yet powerful method is traditionally used to find textual similarities between documents, hence it serves as a strong basis for comparison. Due to the very long tail distribution of the lexical items, a Singular-Value-Decomposition (SVD) (Golub & Reinsch, 1971) is used to represent communities in a reduced and denser dimension (other techniques, e.g., PCA and LDA, yielded very similar results). In this research, we used a vector size of 100.

Table 5
Correlation matrix between the different models. Value pairs indicate the Pearson correlation (first value) and Spearman correlation (second value).

	C3	Com2Vec	Doc2Vec	TF–IDF
C3	1.0	0.28 ; 0.4	0.11 ; 0.18	0.1 ; 0.11
Com2Vec		1.0	0.36 ; 0.32	0.15 ; 0.08
Doc2Vec			1.0	0.28 ; 0.2
TF–IDF				1.0

Table 6
Mean error rate of the Top1 and the Top5 measures in the Self-Similarity experiment.

Model	Error rate, Top1	Error rate, Top5
Com2Vec	27.1%	18.3%
TF–IDF	0.78%	0.46%
Doc2vec	0.19%	0.13%
C3	0.13%	0.13%

Paragraph vector (Doc2Vec). The Doc2Vec algorithm learns compact distributed representations of long paragraphs or documents (Le & Mikolov, 2014). The algorithm accounts for the document structure, rather than taking a naive bag-of-words approach. Given that communities are topical and posts (submissions and comments) are structured, we use doc2vec to represent the textual content of each community, where all communities are represented in the same embedding space. We used a window size of 3 and a negative sampling of 5 in creating the vectors.

User-based community embeddings. The Com2Vec algorithm, introduced by Martin (2017), was explicitly designed to capture similarities between communities. Ignoring content altogether, it recovers the membership co-occurrence matrix for each pair of communities, then applies GloVe (Pennington et al., 2014) on the co-occurrence matrix, producing fixed-length embedding vectors. The similarity between communities is computed in that user-embedding space using cosine or any equivalent metric. We use the same hyper-parameter values suggested by Martin (2017).

5.2. Model correlation results

Pearson and Spearman correlations between C3 and the three competitive baselines are reported in Table 5. Correlations were calculated over the distances between all community pairs (1.22M pairs in total). While a positive correlation was found between all models, the correlation values suggest that models may offer different perspectives on the similarity between communities.

5.3. Self-similarity results

In this experimental setting we measure the level of *'self-similarity'* as a sanity check. Dividing each community c_i to two disjoint sub-communities $\overline{c_i}$ and $\widehat{c_i}$, such that $\overline{c_i} \cup \widehat{c_i} = c_i$. We expect the similarity between $\overline{c_i}$ and $\widehat{c_i}$ to be higher than the similarity between c_i and c_j ($i \neq j$). If corresponding sub-communities $\overline{c_i}$ and $\widehat{c_i}$ are not found in the top k communities closest to each other, we consider it an error. We report the mean error rate for two k values: Top1 — in which the counterparts are expected to be the closest to each other, and Top5 — in which we allow the community's counterparts to be among the five closest communities. The results presented in Table 6 demonstrate the competitiveness of C3, achieving the lowest error rate in both k values. Note that both Doc2vec and C3 significantly outperform the TF–IDF and the dedicated Com2Vec models.

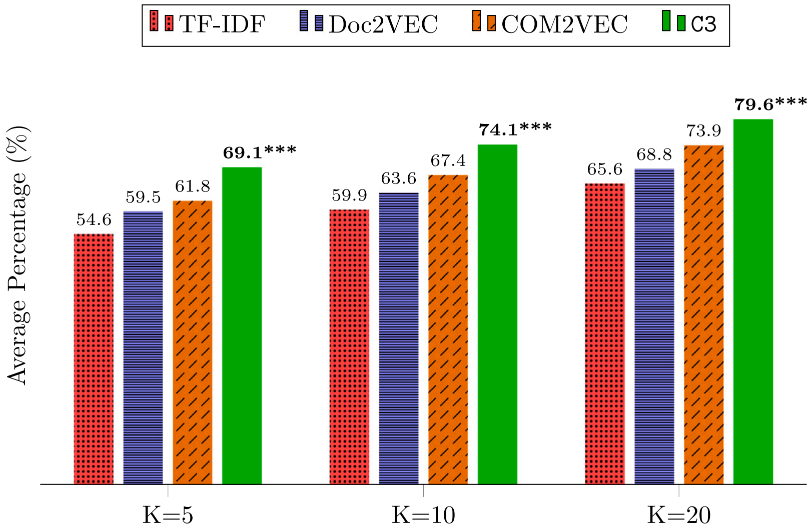


Fig. 2. Average percentage of non-trivial similarities per model. The percentage of non-trivial similarities is calculated for each community $c \in C$ (per each Top-K setting). The y-axis values are the aggregated average over all communities in C. Three different Top-K settings are presented on the x-axis. Best result per Top-K setting is highlighted and statistical significance of the best result is indicated by stars (*** indicates a P-Value $< 10^{-8}$).

Table 7
Evaluation through the annotation task.

	Distinct pairs	Total annotations	Avg. score	Std. score
C3 (Type I)	196	589	3.85	0.58
C3 (Type II)	431	1377	2.56	0.59
Random (Type III)	222	731	1.57	0.38

5.4. Recovering non-trivial similarities

The main premise behind C3 is its ability to discover non-trivial similarities between communities. Given the gold standard categorical hierarchy (see Section 3 and Table 2), we consider a Top K similarity between communities c_i and c_j to be *trivial* if c_i and c_j belong to the same sub-category, and *non-trivial* otherwise. Indeed, C3 recovers a significantly higher percentage of non-trivial similarities, compared to the other methods (Fig. 2). This performance is consistent across all Top-k values.

Non-trivial pairings could be the noisy result of random assignments, thus further validation is required in order to verify the quality of the discovered similarities that are found by C3. One way to validate the results is by having human annotators rank similarities between pairs of communities, then check how these rankings are distributed across pairs of three types: (I) pairs found similar by C3 and are considered trivial, (II) pairs found similar by C3 and are considered non-trivial, and (III) pairs that are matched at random, regardless the similarity level found by C3. An average similarity rank of Type II (non-trivial similarities) that is significantly higher than the similarity rank of Type III (random pairs) demonstrates that the non-trivial similarities recovered by C3 are aligned with a human judgment that is based on a careful and independent study of the communities.

The annotation task was designed as follows: 849 community pairs were sampled at random, 196 of which are of Type I; 431 of Type II; and 222 of Type III. Each annotator was presented with a random sample of 32 community pairs and each pair of the 849 pairs was annotated by (at least) three annotators. Annotators were asked to browse the subreddits of each pair and assign a similarity score on a 1–5 scale. The annotators were not aware of these three pair-types, nor of the gold-standard category and sub-category labels. In order to control and manage the annotation task, we designed and implemented a dedicated web platform.[4] The annotators used the platform for the actual ranking process. The annotation platform also contains detailed instructions for annotators, including examples of communities that maintain a high/low similarity. Two screenshots from the web platform are provided in Fig. 3.

Annotators presented a high agreement, achieving Cohen's kappa of 0.702. We make this annotated corpus available in the project's repository[5] – the first public annotated corpus that contains data that is tagged by the level of similarity between online communities.

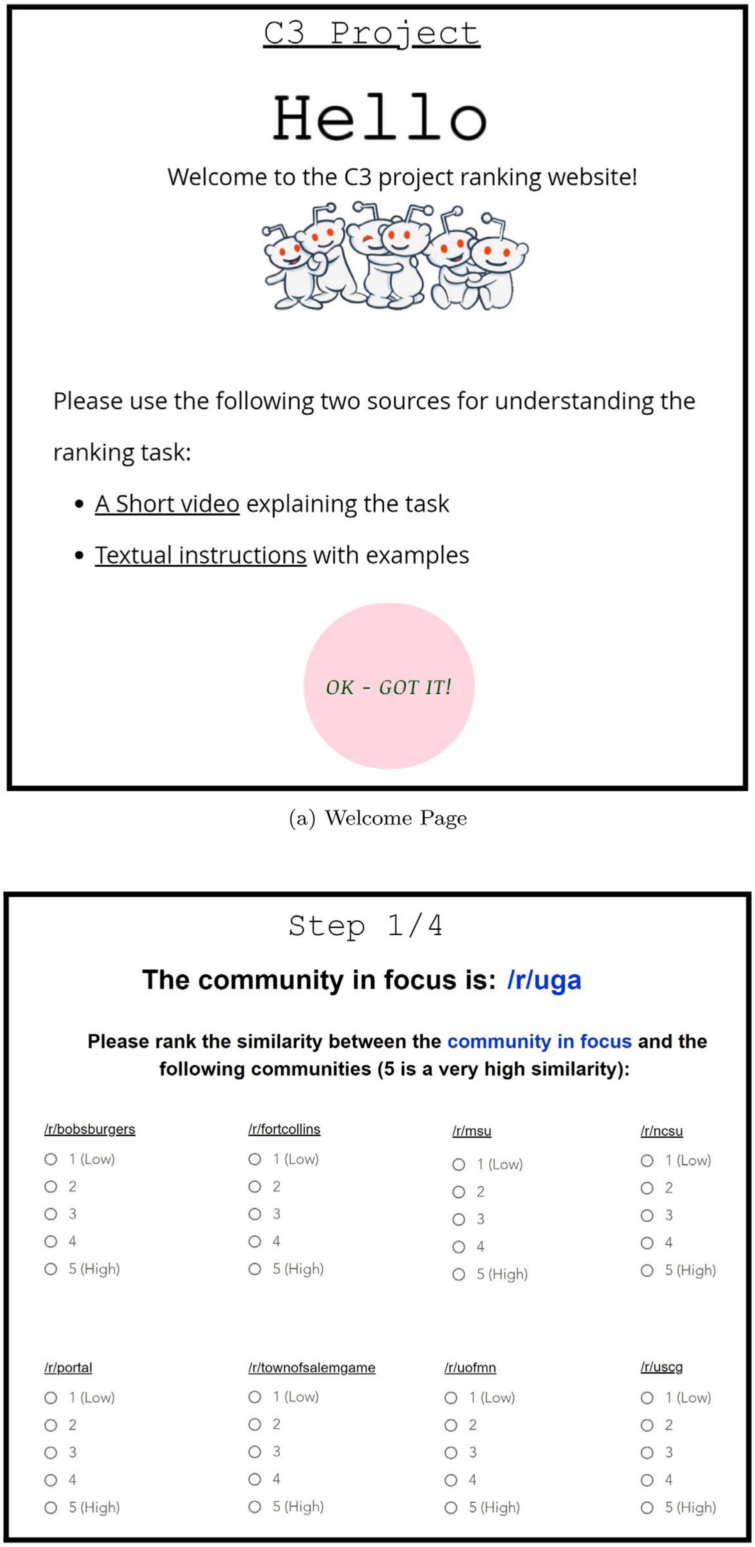
The results of the evaluation through the annotation task are presented in Table 7. Breaking the similarity scores assigned by the annotators to the different pair types, we have an average (/std.) of 3.85(/0.58), 2.56(/0.59), and 1.57(/0.38) for types I, II, and III, respectively. Naturally, the average rank of Type I pairs is the highest (3.85), setting an upper bound for the average similarity rank. We find these differences to be statistically significant (p-value $< 10^{-4}$) using the U-test (Mann & Whitney, 1947). The results over the annotation task highlight that C3 is capable to reveal relevant and meaningful similarities between non-trivial pairs of communities. These results re-validate the benefits of using C3 for an exploratory analysis.

5.5. Sensitivity to the representation's dimension

In this section we explore the impact of varying the dimension of the embedding vectors. This analysis sheds light on the interplay between the various components of our model — the vocabulary alignment, word pairing, and weighting. Specifically, We evaluate the robustness of C3 and other baseline models by controlling the *embeddings vector size*, $|E|$. Per method, we build four contracted models with smaller vector size ($|E| = e \in 10, 50, 100, 200$), and compare each contracted model to the original model ($|E| = 300$). We note that for models based on distributional semantics e is the dimension of the embedding vectors, while for other methods it is the dimension of the feature vector space. The comparison between models and dimensions is measured by the 'Agreement-Level' — the number of communities among the K most similar found by the contracted model that were also among the top K communities found by the original model. We normalize this quantity by K in order to compare across various K values (smaller K values reflect a more conservative comparison).

[4] https://isabrah.wixsite.com/my-site-1.

[5] https://github.com/NasLabBgu/C3-contextual-community-comparison.



C3 Project

Hello

Welcome to the C3 project ranking website!

Please use the following two sources for understanding the ranking task:

- A Short video explaining the task
- Textual instructions with examples

(a) Welcome Page

Step 1/4

The community in focus is: /r/uga

Please rank the similarity between the community in focus and the following communities (5 is a very high similarity):

/r/bobsburgers	/r/fortcollins	/r/msu	/r/ncsu
○ 1 (Low)	○ 1 (Low)	○ 1 (Low)	○ 1 (Low)
○ 2	○ 2	○ 2	○ 2
○ 3	○ 3	○ 3	○ 3
○ 4	○ 4	○ 4	○ 4
○ 5 (High)	○ 5 (High)	○ 5 (High)	○ 5 (High)

/r/portal	/r/townofsalemgame	/r/uofmn	/r/uscg
○ 1 (Low)	○ 1 (Low)	○ 1 (Low)	○ 1 (Low)
○ 2	○ 2	○ 2	○ 2
○ 3	○ 3	○ 3	○ 3
○ 4	○ 4	○ 4	○ 4
○ 5 (High)	○ 5 (High)	○ 5 (High)	○ 5 (High)

(b) Ranking Page (Example)

Fig. 3. Two screen-shots from the website we designed for the human annotation task.

Formally, the agreement level is given in (1), where $C_i^K(m)$ denotes the set of K communities that were found by model m to be the closest to community c_i.

$$AgreementLevel(c_i, e, K) = \frac{C_i^K(|E| = 300) \cap C_i^K(|E| = e)}{K} \tag{1}$$

The results of the robustness analysis are presented in Table 8. The 'Agreement-Level' is presented as the average over all 1565 communities in C. As observed, C3 outperforms other alternative models over all embeddings vector sizes (e) and over all $topK$ values tested. We find this superiority of C3, compared to the alternative models, statistically

Table 8
Sensitivity to the representation's dimension. Correlation and Agreement-Levels of each model with respect to the contracted models over the top K most similar communities. Corr. is the correlation between the original model and the contracted model, over *all* community pairs. The agreement level for each K and model is the average over the whole community corpus ($|C| = 1565$).

			Average agreement level			
	Model	Corr.	$K = 10$	$K = 20$	$K = 50$	$K = 100$
$\vert E\vert = 10$	Com2Vec	0.663	34.1 ± 21.8	36.7 ± 19.9	40.7 ± 18.6	44.9 ± 17.7
	TF–IDF	0.398	9.9 ± 13.0	11.7 ± 12.6	14.3 ± 12.1	16.6 ± 10.7
	Doc2vec	0.109	21.4 ± 18.2	20.7 ± 14.9	17.2 ± 9.5	16.5 ± 6.4
	C3	**0.837**	**57.2 ± 18.8*****	**60.2 ± 16.8*****	**64.6 ± 14.0*****	**67.5 ± 11.6*****
$\vert E\vert = 50$	Com2Vec	0.724	55.7 ± 25.7	55.1 ± 24.2	55.1 ± 23.5	56.5 ± 22.2
	TF–IDF	0.513	30.3 ± 21.2	34.6 ± 19.1	39.9 ± 17.7	43.9 ± 14.9
	Doc2vec	0.355	59.4 ± 24.0	49.7 ± 19.9	35.4 ± 12.4	30.6 ± 8.7
	C3	**0.98**	**81.8 ± 11.9*****	**83.4 ± 9.8*****	**85.9 ± 6.6*****	**87.3 ± 5.4*****
$\vert E\vert = 100$	Com2Vec	0.747	60.4 ± 27.4	59.6 ± 26.5	59.1 ± 25.4	60.0 ± 24.2
	TF–IDF	0.742	43.4 ± 21.5	48.1 ± 19.5	52.2 ± 15.4	55.1 ± 12.2
	Doc2vec	0.47	71.6 ± 20.6	61.2 ± 18.8	45.1 ± 13.2	38.8 ± 9.6
	C3	**0.99**	**84.9 ± 11.2*****	**86.3 ± 8.7*****	**88.6 ± 5.9*****	**90.0 ± 4.6*****
$\vert E\vert = 200$	Com2Vec	0.745	61.6 ± 29.3	60.4 ± 28.3	59.5 ± 27.3	60.3 ± 26.2
	TF–IDF	0.895	66.7 ± 21.3	70.0 ± 18.8	72.5 ± 14.6	74.5 ± 11.3
	Doc2vec	0.578	78.4 ± 17.2	70.4 ± 16.7	54.6 ± 13.3	47.3 ± 10.2
	C3	**0.99**	**88.1 ± 9.9*****	**89.4 ± 6.9*****	**91.5 ± 4.6*****	**93.0 ± 3.2*****

Statistical significance of the best result per category is indicated by stars (*** indicates a P-value $< 10^{-8}$).

Table 9
C3 run time statistics (in seconds) on a Intel Xeon(R) Gold 6130 CPU processor with a processor base frequency of 2.10 GHz. Measures refer to the computational execution time of C3 between community pairs. Steps A–F refer to the calculation steps in Stage B of C3.

	Avg.	Median	Std.
Steps A–F, online calculation of step B	41.22	18.49	64.43
Steps A–F, offline calculation of step B	18.54	9.32	26.34

Table 10
A sample of community pairs that are found similar by C3 but not by any of the other models. The numbers in parenthesis (rightmost column) are the minimum rank among other models.

Community pair	C3 rank	Other models avg. rank (/min. rank)
(hunting, h1z1)	2	539.6 (/94)
(adhd, opiates)	3	139.3 (/68)
(theDonald, leftWithoutEdge)	3	157.6 (/54)
(fantasyhockey, fantasypl)	8	283.3 (/115)
(pokemonGoLA, konosuba)	7	1223.3 (/582)
(singing, newTubers)	6	591.0 (/493)
(socialism, askTrumpSupporters)	6	276.0 (/65)
(tattoos, dirtyr4r)	6	642.3 (/326)
(gay, socialAnxiety)	7	165.3 (/71)
(ukPolitics, australia)	8	296.6 (/69)

significant (p-value $< 10^{-8}$) using the U-test (Mann & Whitney, 1947). We elaborate more on this experiment and its results in Section 6.

C3 running time. The running times of the different steps of the algorithm are presented in Table 9. The recorded times were achieved using an Intel Xeon(R) Gold 6130 CPU processor with a processor base frequency of 2.10 GHz. We notice that Step B in Stage 2 of C3 (see Section 4.2 and Fig. 1) takes a relative high portion of the execution time. However, this specific step can be done 'offline' per community. Hence, we calculated the execution time with and without this specific step. Notice that C3 is independently calculated per pair of communities — facilitating a straightforward parallelization. We note that run times could be further improved with a minimal loss by using a lower embedding dimension. We refer the reader to Sections Section 5.5 and the complement discussion in Section 6 for an analysis of the impact of the embedding space.

5.6. Qualitative analysis

We complement the quantitative results with qualitative analysis from two perspectives: (i) A sample of community pairs that were found similar by C3 but not by *any* of the other models, and (ii) A careful exploration of some of the ten closest communities, found by each of the models, to *two* exemplary community – *babyBumps* and *wallStreetBets*. These qualitative inspections demonstrate, yet again, the power of C3 in recovering unique nuanced similarities.

Uniqueness of C3. Table 10 presents community pairs that were found similar by C3 (among the Top 10 most similar), but were not found similar by *any* of the other models. For example, C3 finds *r/opiates* to be the 3rd closest community to *r/adhd*, while the average similarity rank assigned to that pair by the other models is 139.3. Similarly, C3 finds *r/socialAnxiety* to be the 7th closest community to *r/gay*, while the average similarity rank assigned to this pair by the other models is 165.3. Interestingly, similarities are found between seemingly polarized community pairs, e.g., *r/socialism* and *r/askTrumpSupporters*. These similarities, recovered due to the way C3 balances content, community structure, and patterns of user engagement, provide the social science research community with unique insights in an exploratory manner.

To the bump and beyond. The *babyBumps* community is defined as *"A place for pregnant redditors, those who have been pregnant, those who wish to be in the future, and anyone who supports them"*. Exploring some of the communities found to be closest to *babyBumps* by each of the models is illuminating — yet again demonstrating the strength of C3 in finding interesting and non-trivial similarities. Six closest communities to *babyBumps* out of the top-10 are provided in Table 11 (top), along with the similarity rank assigned to each community by other models. For example, the *beyondTheBump* community (dedicated to new parents) is found similar to *babyBumps* by all models. On the other hand, the similarity of the *weddingPlanning* community to *babyBumps*, ranked #4 by C3, is ranked #3, #752, and #587 by Com2Vec, Doc2Vec, and TF–IDF, respectively. While the similarity between a community for seeking advice on wedding planning and a community supporting expecting parents is intuitively obvious, this similarity is overlooked by two of the other models. Similarly, we intuitively expect to find *babyBumps* and *depression* communities similar, especially given the prevalence of postpartum depression, but this similarity is best recovered by C3.

Examining other communities C3 finds closest to *babyBumps*, suggests that C3 captures the non-trivial similarities between communities providing emotional support, even though each community may be focused on very different challenges and needs, e.g., womanhood

Table 11
The top-10 communities found closest to the babyBumps (top) and wallStreetBets (bottom) communities by each model. Numbers in parenthesis indicate the rank of the community found by the other models. Since $|C| = 1565$, rank values lie on the 1–1564 range.

C3†	Com2Vec‡	Doc2Vec§	TF–IDF$^{\$}$
beyondTheBump ($1^{\ddagger}$, $11^{\S}$, $1^{\$}$)	beyondTheBump ($1^{\dagger}$)	adhd ($15^{\dagger}$)	beyondTheBump ($1^{\dagger}$)
twoXChromosomes ($7^{\ddagger}$, $368^{\S}$, $4^{\$}$)	crochet ($489^{\dagger}$)	accounting ($255^{\dagger}$)	studentNurse ($815^{\dagger}$)
askWomen ($5^{\ddagger}$,$46^{\S}$,$644^{\$}$)	weddingPlanning ($4^{\dagger}$)	androidApps ($613^{\dagger}$)	birthControl ($338^{\dagger}$)
weddingPlanning ($3^{\ddagger}$, $752^{\S}$, $587^{\$}$)	knitting ($325^{\dagger}$)	aspergers ($173^{\dagger}$)	twoXChromosomes ($2^{\dagger}$)
keto ($35^{\ddagger}$, $68^{\S}$, $585^{\$}$)	askWomen ($3^{\dagger}$)	advancedRunning ($277^{\dagger}$)	cancer ($709^{\dagger}$)
depression ($276^{\ddagger}$, $29^{\S}$, $51^{\$}$)	raisedByNarcissists ($9^{\dagger}$)	alberta ($569^{\dagger}$)	herpes ($654^{\dagger}$)
dogs ($8^{\ddagger}$, $12^{\S}$, $617^{\$}$)	twoXChromosomes ($2^{\dagger}$)	anxiety ($22^{\dagger}$)	bigdickproblems ($565^{\dagger}$)
aww ($45^{\ddagger}$, $41^{\S}$, $142^{\$}$)	dogs ($7^{\dagger}$)	advice ($14^{\dagger}$)	hypothyroidism ($904^{\dagger}$)
raisedbynarcissists ($6^{\ddagger}$, $168^{\S}$, $734^{\$}$)	waltDisneyWorld ($308^{\dagger}$)	buildaPCForMe ($524^{\dagger}$)	konmari ($1361^{\dagger}$)
stopDrinking ($146^{\ddagger}$, $408^{\S}$, $778^{\$}$)	thesims ($384^{\dagger}$)	animeSuggest ($643^{\dagger}$)	prolife ($923^{\dagger}$)
askTrumpSupporters ($14^{\ddagger}$, $840^{\S}$, $702^{\$}$)	pennyStocks ($936^{\dagger}$)	weedStocks ($45^{\dagger}$)	pennyStocks ($936^{\dagger}$)
economics ($7^{\ddagger}$, $26^{\S}$, $29^{\$}$)	weedStocks ($45^{\dagger}$)	pennyStocks ($936^{\dagger}$)	weedStocks ($45^{\dagger}$)
askthe_donald ($16^{\ddagger}$, $199^{\S}$, $370^{\$}$)	accounting ($76^{\dagger}$)	securityAnalysis ($1283^{\dagger}$)	securityAnalysis ($1283^{\dagger}$)
teslaMotors ($10^{\ddagger}$, $16^{\S}$, $612^{\$}$)	poker ($169^{\dagger}$)	dashPay ($1137^{\dagger}$)	forex ($801^{\dagger}$)
bestOf ($35^{\ddagger}$, $1220^{\S}$, $100^{\$}$)	forex ($801^{\dagger}$)	litecoin ($1377^{\dagger}$)	4chan ($24^{\dagger}$)
sandersForPresident ($113^{\ddagger}$, $23^{\S}$, $795^{\$}$)	golf ($194^{\dagger}$)	forex ($801^{\dagger}$)	findareddit ($668^{\dagger}$)
conservative ($61^{\ddagger}$, $35^{\S}$, $472^{\$}$)	economics ($2^{\dagger}$)	monero ($470^{\dagger}$)	accounting ($76^{\dagger}$)
marchAgainstTrump ($23^{\ddagger}$, $44^{\S}$, $1180^{\$}$)	business ($399^{\dagger}$)	ethTrader ($44^{\dagger}$)	rickAndMorty ($168^{\dagger}$)
wayOfTheBern ($127^{\ddagger}$, $12^{\S}$, $684^{\$}$)	the_donald ($19^{\dagger}$)	fulfillmentByAmazon ($744^{\dagger}$)	apocalypseRising ($1051^{\dagger}$)
advice ($292^{\ddagger}$, $410^{\S}$, $303^{\$}$)	teslamotors ($4^{\dagger}$)	poker ($169^{\dagger}$)	memeEconomy ($259^{\dagger}$)

(*askWoman*), addiction (*stopDrinking*) or abusive parents (*raisedByNarcissists*). These nuanced similarities are not well captured by other models.

Markets, politics and cars. The GameStop short squeeze of early 2021, organized and promoted in the *WallStreetBets* subreddit, is argued to have shifted the financial power balance. The meaning of the events initiated a debate among economists and sociologists, trying to understand their causes and their impact on future trade. Looking at the top-10 communities found most similar to *WallStreetBets* by C3 (Table 11, bottom) provides a useful computational tool, supporting the social analysis and expanding its perspective. For example, early analysis by Di Muzio (2021) and Long et al. (2021) point to the disdain for 'Big Finance' and the "economic destruction they have brought on jobs, communities, and families that often coincide with their financial practices" (Di Muzio). Indeed, C3 finds similarities to other communities that promote a similar sentiment toward the financial establishment, e.g., *sandersForPesident* and *wayOfTheBern*. On the other hand C3 offers a wider perspective, finding similarities to right-leaning communities like *askTrumpSupporters*, and other "for-profit" communities. These similarities were not recovered by other models. C3 also suggests a connection to the *teslaMotors* community. This is especially interesting since Elon Musk of Tesla was instrumental in the short squeeze rally. We note that this connection was found based on data that were collected three years *prior* to the GameStop events, again, demonstrating the power of C3.

6. Discussion

Among the three alternative models we compare C3 against, only the Com2Vec model (Martin, 2017) takes into account the identity of the community members. However, it does not support a comparison of communities across different platforms (e.g., a Facebook group and a Reddit community) or communities on platforms that maintain anonymity. Due to the way C3 leverages the community structure with the textual data, it is capable of comparing communities across different platforms even when no information about users' overlap is available.

C3 finds similarities in an unsupervised manner. Moreover, it is designed to study similarities without a predefined specific axis (e.g., loyalty) and is capable of revealing non-trivial similarities better than other models. These are unique capabilities that require only a minimal calibration effort. The similarities recovered by the model may be instrumental for a domain expert or a social scientist interpreting various social phenomena. As such, C3 can serve as an ideal exploration tool for social scientists.

One surprising result is that C3 performs well even when relying on very compact embeddings. For example, the Agreement level of the $|E| = 50$ and the big model is over 81% for all K values and over 88% for $|E| = 200$, see Section 5.5. Moreover, C3 in its compact settings with $|E| \geq 50$ significantly outperforms the other models in their larger setting (e.g., $|E| = 200$). We attribute the robustness of C3 to variations in the embedding dimension to one of its fundamental properties: The semantic comparison between communities is not based on a direct comparison between e_{w,c_i} and e_{w,c_j}. Instead, the comparison is based on the respective distances of pairs of words within each community, as specified in Step B (Section 4.2). The semantic relations between pairs are less sensitive to variations in the embedding size.

Error analysis. To better understand the limitations of C3 we provide a brief error analysis, focusing on the false positive/negative predictions of the model.

To better understand the false-positive cases, we manually analyzed the relevant communities on Reddit. We do find evidence of relevant connections between these pairs of communities. For example, C3 finds the *photography* and the *android* communities to be similar (see Table 12, top). Manual examination of the content reveals that the *android* community deals extensively with specs and functionality of the cameras in mobile devices. This focused interest can explain the false prediction. Similarly, while annotators did not see similarity between the *fulfillmentByAmazon* (*"discussions about selling on Amazon and using their Fulfillment by Amazon (FBA) service"*) and *freelance* communities, a more careful examination shows that the Amazon marketplace attracts many small and independent sellers, practically freelancing in the marketplace. Both communities discuss tax and financial issues and the way these should be handled with respect to the authorities or to other business partners.

Turning to some False-Negative examples (Table 12, bottom), we observe that the annotators (engineering students) intuitively indicate a strong connection between *gamingPC* and *nvidia*, the maker of graphic cards popular by gamers, while the model does not latch on this connection. The similarity between the *littleWitchAcademia* and *blech* communities is explained by human annotators as related to the magna and anime culture[6] – a connection that is missed by the model due to the fact that most of the content shared in these communities is graphic (images, videos) rather than textual. This highlights one limitation of

[6] Comics, graphic, hand-drawn, and computer-generated animations originating from Japan.

Table 12
Error Analysis. False-Positive (FP) and False-Negative (FN) examples of C3. Human Ann. is the average human annotation of the communities pair (spans over [1, 5]). The C3 Ranks column contains two values (spans over [1, 1564]) in order to represent both ranks of the communities pair (e.g., *android* is ranked 2nd in the list of communities most similar to *photography* while *photography* is ranked 11th in the list of communities most similar to *android*).

	Community pairs	Human ann.	C3 distance	C3 ranks
FP	(photography, android)	2.33	0.308	(2, 11)
	(fulfillmentByAmazon, freelance)	2.33	0.526	(6, 1)
	(videos, books)	1.0	0.314	(27, 18)
FN	(gamingPC, nvidia)	4.33	2.25	(986, 1233)
	(androidGaming, n64)	4.67	1.05	(981, 605)
	(littleWitchAcademia, bleach)	4.67	2.37	(876, 1444)

C3 — even big and active communities are not well compared if they do not share enough textual content.

The error analysis also highlights an often neglected aspect related to human annotation, especially with regard to cognitively demanding tasks (Joseph et al., 2017). Visual similarity is easy to detect, while complex financial and taxation issues buried in longer texts are harder to find. Similarly, the relation between specific hardware and gaming is evident to gamers and CS/Engineering students (our annotators), while more opaque to other annotators.

7. Conclusion and future work

We introduce an unsupervised method for a direct Contextual Community Comparison (C3). It recovers unique non-trivial similarities by leveraging both text, community structure, and the roles community members assume. The benefits of the method have been demonstrated quantitatively and qualitatively through an array of experiments. Future work will take two trajectories: (i) Address similarities between communities across different platforms, and (ii) Add an interpretability component for the model. We assume that a short "explanation" together with the model's output would shed light on the reasons behind a model's output. Inspired by the latest interpretability components (Lundberg et al., 2020; Lundberg & Lee, 2017), we plan to study the different features (e.g., tokens, users) that highly contribute to pushing the C3 model to its final similarity score.

CRediT authorship contribution statement

Abraham Israeli: Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing. **Shani Cohen:** Conception and design of study, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing. **Oren Tsur:** Conception and design of study, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We made the required data available as part of the project. The relevant links appear in the manuscript.

Acknowledgment

We thank Dr. Hila Gonen for valuable discussions and advice about this research.

References

Abd Rahman, R., Omar, K., Noah, S. A. M., & Danuri, M. M. (2018). A survey on mental health detection in online social network. *International Journal on Advanced Science, Engineering and Information Technology*, *8*(4–2), 1431.

Artetxe, M., Labaka, G., & Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2289–2294).

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media, vol. 14* (pp. 830–839).

Blodgett, S. L., Green, L., & O'Connor, B. (2016). Demographic dialectal variation in social media: A case study of african-American english. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1119–1130).

Bouarara, H. A. (2021). Recurrent neural network (RNN) to analyse mental behaviour in social media. *International Journal of Software Science and Computational Intelligence (IJSSCI)*, *13*(3), 1–11.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Cheong, P. H., Poon, J. P., Huang, S., & Casas, I. (2009). The internet highway and religious communities: Mapping and contesting spaces in religion-online. *The Information Society*, *25*(5), 291–302.

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on world wide web* (pp. 307–318).

Datta, S., & Adar, E. (2019). Extracting inter-community conflicts in reddit. In *Proceedings of the international AAAI conference on web and social media, vol. 13* (pp. 146–157).

Del Tredici, M., & Fernández, R. (2017). Semantic variation in online communities of practice. In *Proceedings of the 12th international conference on computational semantics*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Di Muzio, T. (2021). *GameStop capitalism. Wall street vs. The reddit rally (Part I)*. Toronto: The Bichler and Nitzan Archives.

Eisenstein, J. (2013). Identifying regional dialects in on-line social media. In *The handbook of dialectology, vol. 2013* (pp. 368–383). Wiley Online Library.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1277–1287).

Fuchs, C. (2021). *Social media: a critical introduction*. SAGE Publications.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(16), E3635–E3644.

Golub, G. H., & Reinsch, C. (1971). Singular value decomposition and least squares solutions. In *Linear algebra, vol. 2* (pp. 134–151). Springer.

Gonen, H., Jawahar, G., Seddah, D., & Goldberg, Y. (2020). Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 538–555).

Hamilton, W., Zhang, J., Danescu-Niculescu-Mizil, C., Jurafsky, D., & Leskovec, J. (2017). Loyalty in online communities. In *Proceedings of the international AAAI conference on web and social media vol. 11, no. 1*.

Hanel, P. H., Maio, G. R., & Manstead, A. S. (2019). A new way to look at the data: Similarities between groups of people are large and important. *Journal of Personality and Social Psychology*, *116*(4), 541–562.

Hessel, J., Tan, C., & Lee, L. (2016). Science, askscience, and badscience: On the coexistence of highly related communities. In *Proceedings of the international AAAI conference on web and social media, vol. 10, no. 1* (pp. 171–180).

Hofstede, G. (2001). *Culture's consequences: comparing values, behaviors, institutions and organizations across nations*. Sage publications.

Hornsey, M. J., & Hogg, M. A. (2000). Intergroup similarity and subgroup relations: Some implications for assimilation. *Personality and Social Psychology Bulletin*, *26*(8), 948–958.

Huffaker, D., Jorgensen, J., Iacobelli, F., Tepper, P., & Cassell, J. (2006). Computational measures for language similarity across time in online communities. In *Proceedings of the analyzing conversations in text and speech* (pp. 15–22).

Janchevski, A., & Gievska, S. (2019). A study of different models for subreddit recommendation based on user-community interaction. In *Proceedings of the international conference on ICT innovations, vol. 1110* (pp. 96–108). Springer.

Joseph, K., Friedland, L., Hobbs, W., Lazer, D., & Tsur, O. (2017). ConStance: Modeling annotation contexts to improve stance classification. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 1115–1124).

Jurgens, D., Tsvetkov, Y., & Jurafsky, D. (2017). Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th annual meeting of the association for computational linguistics (short papers), vol. 2* (pp. 51–57).

Kim, H. K., & McKenry, P. C. (1998). Social networks and support: A comparison of african Americans, Asian Americans, caucasians, and hispanics. *Journal of Comparative Family Studies*, *29*(2), 313–334.

Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). Community interaction and conflict on the web. In *Proceedings of the world wide web conference* (pp. 933–943).

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of international conference on machine learning* (pp. 1188–1196).

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, *4*(10), 1021–1028.

Lin, B., & Vassar, J. A. (2009). Determinants for success in online learning communities. *International Journal of Web Based Communities*, *5*(3), 340–350.

Long, C., Lucey, B. M., & Yarovaya, L. (2021). I just like the stock" versus" fear and loathing on main street": The role of reddit sentiment in the GameStop short squeeze. *SSRN Electronic Journal*, *31*, 1–37.

Lu, A., Wang, W., Bansal, M., Gimpel, K., & Livescu, K. (2015). Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 250–256).

Lucy, L., & Bamman, D. (2021). Characterizing english variation across social media communities with BERT. *Transactions of the Association for Computational Linguistics*, *9*, 538–556.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4768–4777).

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 50–60.

Martin, T. (2017). Community2vec: Vector representations of online communities encode semantic relationships. In *Proceedings of the second workshop on NLP and computational social science* (pp. 27–31).

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

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems, vol. 2* (NIPS), (pp. 3111–3119).

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics*, *42*(3), 537–593.

Nguyen, D., & Rose, C. (2011). Language use as a reflection of socialization in online communities. In *Proceedings of the workshop on language in social media* (pp. 76–85).

Noor, S., Guo, Y., Shah, S. H. H., Nawaz, M. S., & Butt, A. S. (2020). Research synthesis and thematic analysis of twitter through bibliometric analysis. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *16*(3), 88–109.

Olson, R. S., & Neal, Z. P. (2015). Navigating the massive world of reddit: Using backbone networks to map user interests in social media. *PeerJ Computer Science*, *1*, Article e4.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, vol. 1 (long papers)* (pp. 2227–2237).

Sahoo, S. R., & Gupta, B. B. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing*, *100*, Article 106983.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.

Sankoff, G., & Blondeau, H. (2007). Language change across the lifespan:/r/in montreal french. *Language*, *83*(3), 560–588.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *SIGKDD Explorations Newsletter*, *19*(1), 22–36.

Smith, S. L., Turban, D. H., Hamblin, S., & Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv preprint arXiv:1702.03859.

Soon, C., & Kluver, R. (2007). The internet and online political communities in Singapore. *Asian Journal of Communication*, *17*(3), 246–265.

Spertus, E., Sahami, M., & Buyukkokten, O. (2005). Evaluating similarity measures: A large-scale study in the orkut social network. In *Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining* (pp. 678–684).

Stier, S., Posch, L., Bleier, A., & Strohmaier, M. (2017). When populists become popular: Comparing facebook use by the right-wing movement pegida and german political parties. *Information, Communication & Society*, *20*(9), 1365–1388.

Tinto, V., & Love, A. G. (1995). A longitudinal study of learning communities at LaGuardia community college. ERIC.

Tran, T., & Ostendorf, M. (2016). Characterizing the language of online communities and its relation to community reception. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1030–1035).

Van, D., & Johannes, A. G. M. (2012). *The network society* (3rd ed.). SAGE Publications.

Van den Berg, P. T., & Wilderom, C. P. (2004). Defining, measuring, and comparing organisational cultures. *Applied Psychology*, *53*(4), 570–582.

Waller, I., & Anderson, A. (2019). Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms. In *Proceedings of the world wide web conference* (pp. 1954–1964).

Zhang, J., Hamilton, W., Danescu-Niculescu-Mizil, C., Jurafsky, D., & Leskovec, J. (2017). Community identity and user engagement in a multi-community landscape. In *Proceedings of the international AAAI conference on web and social media, vol. 11, no. 1* (pp. 377–386).

Zhang, J., & Sun-Keung Pang, N. (2016). Investigating the development of professional learning communities: Compare schools in Shanghai and Southwest China. *Asia Pacific Journal of Education*, *36*(2), 217–230.

Zhang, L., & Wang, J. (2018). Why highly cited articles are not highly tweeted? A biology case. *Scientometrics*, *117*(1), 495–509.