Community identities under perturbation: COVID-19 and the r/digitalnomad subreddit
by Manuel Pita, Karine Ehn, and Thiago dos Santos

Abstract
Digital nomads (DNs) are hyper-mobile, location-independent workers whose practices blur traditional boundaries between labour, leisure, home and travel. They rely on digital tools to work and on computer-mediated communication to share knowledge and resources. Their resource-sharing culture is vital for self-efficacy and self-actualisation — two fundamental values that define the DN identity. Community identity is a constant social-semiotic construct mutually determined by (micro) interactions and the (macro) influence of collectively shared meanings and symbols. However, most of our understanding of community identity comes from structural and synchronic properties that often assume identity “exists” as an entity, separate from underlying collective dynamics. In this paper, we approach community identity diachronically, by introducing a quantitative typology that projects conversational timelines on two dimensions relevant to understanding the process of community identity construction: (a) the temporal orientation to the community core (or peripheral) conversation topics, and (b) interaction pattern anomalies. We cast three years of r/digitalnomad threads as a set of conversation topics, describing the interaction dynamics on these topics using the proposed typology. The central questions asked by this paper are whether there was a pre-pandemic stable community expression, and if so, how the COVID-19 pandemic may have perturbed it. Since lockdowns and travel restrictions impinged on fundamental DN values, the nature of the topics and interaction patterns that characterise the r/digitalnomad subreddit could have changed its character. We found a stable pre-pandemic balanced expression of core and peripheral conversation topics with regular interaction patterns. This identity expression was perturbed temporarily in the middle of the lockdown period when the community shifted focus to interactions about visa issues. As many countries began to reopen their borders around May 2021, a record-breaking number of interactions disrupted identity expression more profoundly. First, we observed constant interaction anomalies. Second, the community orientation revealed multi-factorial emergent issues, most of which revolved around conversations about what it means to be a DN, resource sharing and restrictions. We hypothesise that an influx of outsiders may have caused a clash of social norms and triggered a transformation of the DN identity that was still ongoing at the end of the studied period, in December 2021.

Contents
Introduction
Methods
Results
Conclusions
Discussion
Introduction

Digital nomads

Makimoto and Manners (1997) foretold the digital nomad (DN) emergence in a globalised world, where the boundaries between work, leisure, home and travel would increasingly disappear. Today, the DN community comprises professionals with skills that allow them to work from almost any place with an Internet connection (see Thompson [2019] for an overview).

Recent research conceptualises DNs as a contemporary counter-culture that emerged in the new scenarios created by globalisation — particularly by the rapid evolution of the hyper-connected information ecosystem (D’Andrea, 2006). However, unlike social movements, e.g., feminism or LGBT rights, DNs do not have a political agenda or a central collective goal. Instead, DNs pursue individual goals using life-hacking techniques to overcome challenges and obstacles (Wang, et al., 2018).

An essential aspect of the DN life-hacking culture is sharing knowledge and resources openly with their peers (Faraj, et al., 2011). In pursuing life goals, DNs appear to prioritise Maslowian self-actualisation (McLeod, 2007) over ensuring basic human needs. Indeed, self-actualisation is a crucial factor that defines the DN identity (Beck-Gernsheim, 2002; D’Andrea, 2006; McLeod, 2007; Lyng, 2008). Several studies consider DNs a critical case study for understanding trends and forecasting future work practices (see Sutherland and Jarrahi, 2017; Hemsley, et al., 2020; de Almeida, et al., 2021; de Almeida, et al., 2022).

Since DNs are highly mobile in the physical world, most of their community practices take place online (see Wang, et al., 2018; Priante, et al., 2018; Thompson, 2019). Public DN conversations take place mainly on Reddit but also on Twitter. In these channels, DNs often engage in question-answer dynamics, similar to other online communities, like StackOverflow — a support community for computer programmers (see Wang, et al., 2013).

Location independence was a vital aspect of the DN lifestyle before the COVID-19 pandemic-imposed lockdowns and remote work as norms for public health (D’Andrea, 2006; Aroles, et al., 2020; Bauman, 2000; Hemsley, et al., 2020; Bozzi, 2020; Barbieri, et al., 2021). One of the immediate consequences of lockdowns was the so-called DN dilemma: the choice to stay abroad without being able to travel for an unforeseeable future vs. finding a way to return home.

In previous work, we analysed content created by DNs with a large following on YouTube at the start of the pandemic (Ehn, et al., 2022). We found that creators contributed to reinforcing the community’s values during a crisis and did not favour alternatives to the DN dilemma. These results were one of the motivations to study how the pandemic affected the expression of the identity of the DN community, by studying their public conversations on social media.

Previous relevant research includes work by Zhang, et al. (2017), who proposed a typology to study community identity based on two synchronic features, namely how distinctive and dynamic a community’s interests are. Zhang and colleagues analysed and compared almost 300 Reddit communities using this typology.

In addition, Hemsley, et al. (2020) studied modern forms of work, physical mobility, and technology practices before the COVID-19 pandemic by analysing topic models of Twitter data. They showed that delineating topical boundaries in the network of language interactions explained salient features of the evolution of the DN community identity via self-reinforcing messages about hyper-mobility and leisure.

Finally, two recent papers used a grounded theory approach to study a collection of Reddit threads in the r/digitalnomad subreddit shared in the first few months after the World Health Organisation announced the COVID-19 pandemic on 11 March 2020 (de Almeida, et al., 2021) and for a further period, until July 2021 (de Almeida, et al., 2022). de Almeida, et al. found that the DN community had an influx of aspiring new community members who joined because of the pandemic, testing the subreddit’s interaction norms.
They also found a significant increase in long DN personal narratives on the pandemic’s effects, and an increase in the number of conversations about border control and dealing with visa expiration during lockdowns.

In this work, we analyse threads in the r/digitalnomad subreddit collected between 1 January 2019 and 31 December 2021. Specifically, we propose a community identity typology based on two dimensions that capture the dynamics of the community’s social-semiotic identity construction process.

The first dimension of the typology quantifies the degree to which interactions (in relatively short time intervals) are oriented towards the community core conversations, or shift to other (peripheral) topics instead. The second dimension measures the degree to which the interaction pattern in some time interval is “unexpected”, i.e., anomalous, in the context of all other patterns of other time intervals.

In broad terms, this paper begins by asking whether the r/digitalnomad community displayed interactional regularities before the COVID-19 pandemic and how we might use these regularities to define baseline collective identity traits. Then, the paper asks whether, and how, the pandemic affected the r/digitalnomad community identity, provided there was a stable pre-pandemic baseline community identity.

Reddit

Reddit is a leading online social network platform with over 50 million daily active users. Two features that distinguish it from other platforms are (a) their focus on topical communities called subreddits (most of which are public), and (b) subreddits are conversational by design. Reddit imposes very few restrictions on conversation topics, post length, use of different media types and nested conversational structure.

Subreddits are (unofficially) moderated by community volunteers. One of the main goals of moderation is to avoid existing conversations to be restarted in new threads. When users try to post a new thread on a subject that already exist in the subreddit, moderators will not approve it, which encourages redditors to search before they attempt to post (see Proferes, et al. [2021] for a comprehensive review).

A subreddit contains threads on conversation topics that are relevant to a specific online community. A thread is largely defined by its original content, henceforth [OC] (using Reddit’s jargon, which also includes the shorthand OP to refer to the original poster). [OC] have a title and a body. People can post comments to the [OC], reply to those comments or post replies to a reply. Redditors can upvote or downvote any post in a thread. The Reddit algorithm uses voting information to increase or decrease thread visibility on the subreddit’s front page. People may visit a subreddit’s site to see what content is trending, but the most common driver of interaction is the search for specific conversations.

While anyone can read subreddits online, only people with a Reddit account can post and interact. However, account creation does not enforce identity verification. Many redditors value this anonymity, but the mechanism encourages people to interact under different pseudonyms (Ammari, et al., 2019).

Reddit allows data collection for research purposes, as long researchers use the free official API (Reddit.com, 2021). However, moderators of specific subreddits may impose additional restrictions on using data for research (see Proferes, et al., 2021).

Community identity

Identity is a multifaceted phenomenon investigated in various disciplines, mainly sociology and social psychology. In the former, identity theory studies how individuals relate to shared meanings and norms, enacting and adjusting them through interactions with others. Social identity theory (social psychology), on the other hand, concentrates on how social groups realise their identity through intra- and inter-group dynamics (see Davis, et al. [2019] and references therein).
Most research on identity has adopted a functionalist (or essentialist) approach that assumes the existence of identity as an enduring entity, subject to empirical enquiry. One of the methodological lines of functionalist research on community identity relies on the study of the structural properties of social networks (Colombo and Senatore, 2005). Since such a structural approach does not account for dynamics, claims about identity come from topological metrics describing how people are connected to each other, and not from processes through which shared meanings are constantly created and re-interpreted. Indeed, scholars have debated on the structure vs. dynamics perspectives for studying most types of networked phenomena for some time (see Mitchell, 2006).

In sharp opposition to the functionalist approach, we conceptualise community identity as the outcome of networked communication dynamics (Colombo and Senatore, 2005; Benwell and Stokoe, 2006; Gooding, et al., 2007). Previous research has shown that linguistic variation and differential interaction patterns reveal important aspects of a community’s identity (see Cassell and Tversky, 2005; Danescu-Niculescu-Mizil, et al., 2013). In this view, collectively shared meanings are constantly created and revised through interaction.

We approach online communities as complex systems (see Mitchell, 2009, 2006). To understand their identity we seek to (a) identify what emerges from collective dynamics, (b) how what emerges is dynamically expressed with regularities that can be mapped to an identity, and (c) how that identity expression responds to perturbations.

Note that in this work the term collective refers to variables and behaviours that emerge from interactions in an online community, measurable in social-network dynamics. There is no connection with notions such as collective action used to study identity in social movements (as in Melucci, 1996).

Community identity from conversation topics

To identify what emerges in the r/digitalnomad community, we cast the high-dimensional space of their conversation threads into a low-dimensional set of conversation topics.

A conversation topic is a coherent subject that is separable from other subjects. In an extreme example, conversations about “quantum physics” are different from those about “fashion” because it is highly probable that conceptual dimensions relevant to those topics do not overlap. However, natural human conversations mix diverse elementary topics corresponding to single concepts or categories, particularly those in a community with a collective identity.

For instance, consider a Reddit community interested in food. Some of their threads may be specifically about “winter foods” — this is a conversation topic. Different subsets of these threads are likely to focus on sub-topics like seasonal ingredients, winter dessert recipes and cooking procedures (further subdivided into oven settings, cooking times and other categories). An elementary topic is a sub-topic defined by a set of keywords describing a category that cannot be further decomposed.

In this work, we rely on the computation of elementary topics by finding word co-occurrence patterns in a collection of threads from the r/digitalnomad subreddit. Naturally, we can represent a conversation’s topic structure as a hierarchy of increasingly more general topics from (non-decomposable) elementary topics. Furthermore, we can choose to focus on different (coarse or fine-grained) levels of a topic hierarchy for analysis. See Jelodar, et al. (2019) for a recent review on topic modelling and Methods for further details about the chosen topic modelling algorithm. In the remainder of this paper, the term “topic” is a shorthand for “conversation topic”.

A diachronic typology of community identity

To understand how what emerges is dynamically expressed over time, we adopt an interactional framework concerned with the social-semiotic process of identity construction (De Fina, 2015). First, we use community dynamic interaction patterns across topics over time as a proxy to understand the dynamics of identity construction. Then, we represent and describe community identity in terms of the regularities and
anomalies found in those interaction patterns.

This approach does not deny that intersubjective processes underlie the emergence of community identity. However, we are only concerned with the outputs of those processes, i.e., symbols, language and interactions, and not with reasoning about internal cognitive states and mechanisms inside the minds of community members (see Fuhse [2015] for a sociological perspective).

We propose a two-dimensional, diachronic typology to study community identity. The first dimension, Core-Peripheral Orientation, $O(t)$, quantifies the degree to which the community leans to core conversations or drifts to peripheral topics in a given time interval, $t$. Orientation lies in the interval $[1,1]$. Positive (negative) values mean the community is oriented toward core (peripheral) topics during a time interval, $t$. Orientation values close to or equal to zero thus indicate an equal interaction split between core and peripheral topics in the interval.

The second dimension, Anomaly score, $A(t)$, quantifies the degree to which the distribution of community interactions in a time interval, $t$, is expected, i.e., normal, given the community interaction patterns over a set of time intervals. Anomaly values are also in the interval $[1,1]$. Positive (negative) values mean the interactions during, $t$, are anomalous (normal). See Methods for relevant mathematical details and algorithms.

By projecting time intervals on the typology, we can make inferences about how a community responds to perturbation. For example, at a broad level, given the date of an event with a potential impact on the community, we can compare community orientation and anomalous interaction patterns before and after the event. Exploratory analysis can also reveal inflexions in orientation, anomaly scores or both.

This typology is diachronic because it allows inferences about dynamic processes affecting community identity. Even though our typology shares similarities with the work by Zhang, et al. (2017), notice that the latter adopted a synchronic approach, in which temporal information was averaged and mapped onto a descriptive measure of overall “dynamicity” of Reddit communities.

Explaining anomalies

Orientation and anomaly scores describe properties of interaction distributions across topics in specific time intervals. In some cases, knowing what interactions on specific topics contributed the most to classifying a time interval as anomalous might be helpful. In addition, we may also want to know whether there were interactions in other topics that were “pushing” a given time interval toward normality.

For example, suppose there was an isolated anomalous interval that maintained core orientation. The anomaly may have resulted from unexpectedly high interactions in one of the core topics and unexpectedly low interactions in a peripheral topic. Identifying these topics is necessary for interpreting the isolated anomaly. Another example would be observing a sequence of anomalous time intervals. The interactions driving those anomalies may have been on a single topic, meaning there is a common explanation for them. However, it is also possible that anomalies are caused by interactions on different topics, which in more extreme cases, may signal responses to significant perturbations, or that the community may be undergoing an identity transformation.

Methods

Data collection and corpus constitution

The collected raw data comprises 27,951 Reddit threads started by 11,581 unique OPs (Original Poster, see
Introduction) from the r/digitalnomad subreddit, in the period starting on 1 January 2019 and ending on 31 December 2021. The data collection procedure integrated the native Reddit API (https://www.reddit.com/dev/api/) and an external Reddit archive service provided by the PushShift API (https://pushshift.io). PushShift was used to obtain the thread IDs in the chosen date range, and the native API was used to retrieve the content of the conversation threads. The data collection procedure followed all Reddit and the r/digitalnomad subreddit rules (described in the subreddit page). For each post, i.e., [OC], comment or reply the following attributes were kept: author, date and time of posting, title and selftext (body).

A corpus is a collection of text documents, and the set of unique words they contain constitutes the corpus vocabulary. In this work, the method constituted a corpus of threads from the r/digitalnomad subreddit. The method implemented a corpus constitution procedure to clean, sort and select the most appropriate data for the computational and statistical analyses presented in Results. The procedure first created a copy of the collected threads, considering their [OC] only, and disregarding threads with markdown text that indicated they were either deleted (by the OP) or removed (by moderators).

This procedure reduced the corpus to 16,005 threads. For the threads represented in this filtered version of the corpus, the procedure computed (a) the [OC] length, which is the number of words in the string containing both title and body, (b) the thread’s total number of thread interactions, and (c) the thread’s maximum reply depth.

The method arbitrarily chose threads that satisfied two criteria, (a) the [OC] length was at least 10 words, and (b) the total number of thread interactions (sum of comments and replies) was equal to or greater than six. The threshold values for both criteria were within the interquartile range for the corresponding post length and count of thread interactions. Criterion (a) was essential to constitute a corpus with the minimum statistical properties for deriving a topic model (explained later in this section). For criterion (b), the chosen threshold value restricts the data to threads that elicited at least some activity in terms of comments and replies. The final corpus contains 6,029 threads. See Table 1 for summary descriptive statistics.

<table>
<thead>
<tr>
<th>DN_Reddit Corpus</th>
<th>Post length (number of words)</th>
<th>Thread interactions</th>
<th>Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>86</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>Max</td>
<td>3,113</td>
<td>591</td>
<td>86</td>
</tr>
<tr>
<td>IQR (Q1-Q3)</td>
<td>45–150</td>
<td>9–30</td>
<td>3–7</td>
</tr>
</tbody>
</table>

The method leveraged the redditcleaner Python package to remove special markdown characters used by Reddit internally. In addition, stop words, punctuation, special symbols and numbers were removed using the Python NLTK package. Finally, the words in every document were lemmatised using the Python Spacy package. Lemmatisation reduces inflectional and derivationally related linguistic forms of a word to a common linguistic elementary form. Mapping counts of word variants to a single token is critical for text analysis methods that rely on word frequencies, such as the one used here to produce topic models.
Data collection and corpus constitution

The topic modelling method

The method relied on a non-parametric topic-modelling algorithm (Gerlach, et al., 2018) that casts the input corpus as a bipartite network where one type of node represents the input texts (corpus), and the other type represents words from the corpus vocabulary. The algorithm identifies the best modular partition of the network based on stochastic block modelling (see Karrer and Newman [2011] for an introduction). Furthermore, the modular structure is also hierarchical, which allows the analyst to choose the granularity of the output representation.

The chosen topic modelling algorithm has two key advantages: (a) it does not require the pre-specification of the number of topics, and (b) topics are network modules. There is no assumption that topics have idealised distributions of word frequencies.

The method restricted the corpus to [OC] only. This choice ensured that conversation threads were classified into the conversation topic intended by the OP. This methodological choice rests on the reasonable assumption that the threads remain relatively close to the topic defined by the [OC], even though comments and replies may branch out into potentially many sub-topics.

One of the model’s components is a hierarchy of elementary topics, each a mixture of words. The lowest level of the hierarchy contains word partitions that co-occur in the corpus. These partitions are candidate elementary topics. Words in an elementary topic have a relevance score. Candidate elementary topics are validated after inspection. Sometimes a candidate elementary topic contains words that co-occur in the entire corpus — such as most verbs. Such candidate word partitions are not useful elementary topics and are thus disregarded.

In addition, the algorithm produces modules in a hierarchical structure. Thus, higher levels of the topic model hierarchy result from combining similar elementary topics at lower levels. Naturally, elementary topics at low levels are fine-grained (specific) and, conversely, more coarse-grained (general) at higher levels.

Since the algorithm is stochastic (starts with a random topic model and iteratively fine tunes it), computing several models from different initial configurations is important to ensure that the resulting topic model structure is not an artefact. Therefore, the method used five different random seeds. For each random seed, the algorithm was executed from five different random initial conditions, thus producing 25 topic models. Inspection of the resulting models confirmed that there was almost no topic structure variation. The method selected the most modularly compact model, based on a measure of minimal description length computed by the algorithm.

Analysis of elementary and conversation topics

Even though this paper is primarily concerned with interaction patterns, a general description of conversation topics was necessary for interpretation and contextualisation. To describe conversation topics, the method:

1. Iterated over each separately
2. Drew random samples of \(k = 5\) threads (without replacement)
3. Accumulated the frequency distributions of all the words across elementary topics and normalised the result

This step is equivalent to computing the average elementary topic distribution for the conversation topic cluster. The method repeated this process until there were no more threads to sample from or when the normalised accumulated elementary-topic empirical distribution had not changed over at least 10 consecutive iterations. Then, the resulting distribution of elementary topics was used to describe and label
the corresponding conversation subject.

**The Orientation-Anomaly (OA) typology**

**Definitions and math notation**

Assume that a topic model has been derived for a subreddit corpus collected in some period. Let, \( \mathbf{x} \), denote the corresponding vector of topic labels and, \( n \), the vector cardinality, \( |\mathbf{x}| \). To analyse interaction dynamics, the method considered a vector, \( \mathbf{t} \), of equally sized intervals (e.g., days, weeks or months). The vectors \( \mathbf{c}_{[i,t]} = [c_{[x_1,t]}, \ldots, c_{[x_n,t]}] \) contain the interaction counts for all topics, \( x_i \in \mathbf{x} \), in all time intervals, \( t_j \in \mathbf{t} \) and collection of vectors, \( [\mathbf{c}_{[i,t]}] \) for all time intervals, \( t_j \in \mathbf{t} \), constitute the columns of interaction matrix, \( \mathbf{C} \), where each matrix element, \( c_{[i,j]} \), is the count of interactions topic, \( x_j \), in time interval, \( t_j \).

**Core-Peripheral Orientation**

The first dimension included in the proposed OA typology is the core-periphery orientation. This dimension quantifies the ratio of interactions on topics that are central (core) to the community to other, peripheral, topics.

To label topics, \( x_i \in \mathbf{x} \), as core or peripheral, the method quantified the Discrepancy, \( D_{(x_i,t)} \), between the interactions on topic, \( x_j \), in \( t_j \) relative to interactions on \( x_i \), for all other intervals, \( t \in \mathbf{t} \) as their point-wise mutual information, PMI (see Cover and Thomas, 1991),

\[
D(x_i, t) = \log \frac{P(x_i, t)}{P(x_i, t)}
\]

A topic, \( x_i \), with median discrepancy, \( D_{(x_i)} \approx 0 \), is considered a core topic because, on average, the observation of its interaction pattern in one time interval is not “surprising” in the context of all time intervals. A stringent threshold, \( |D_{(x_i)}| \leq 0.05 \), determined the classification of a topic as core. The community’s core-peripheral orientation in time interval, \( t \), is given by,

\[
\mathcal{O}(t) = \frac{\sum_{i=1}^{c'} c'_{[i,t]} - \sum_{j=1}^{c''} c''_{[j,t]}}{\sum_{k=1}^{n} c_{[k,t]}}
\]

where \( c'_{[*,t]} \) denotes the column vector for interval \( t \), but restricted to core topics, and similarly \( c''_{[*,t]} \) contains the counts for peripheral topics in, \( t \). The cardinalities \( |c'| \) and \( |c''| \) correspond to the number of core and peripheral topics, respectively and thus \( |c'| + |c''| = n \).

A time interval, \( t \), with the same number of interactions on core and peripheral topics has orientation, \( \mathcal{O}(t) \). Positive (negative) values indicate core (peripheral) orientation. In the extreme cases,
Community identities under perturbation: COVID-19 and the r/digitalnomad subreddit

\( \mathcal{O}(t) = 0 \)
\( \mathcal{O}(t) = 1 \) means the community only interacted on core topics, and conversely, \( \mathcal{O}(t) = -1 \) means that all interactions were on peripheral topics.

Interaction anomalies

The second dimension in the OA typology scores interaction patterns in specific time intervals on a normal-anomaly scale. Interaction anomalies result from,

1. Unexpected interaction shifts from a topic, \( x_i \) to another topic \( x_j \)
2. The observation of extreme interaction counts, above or below, expected levels for some topic (outliers)
3. Interaction patterns coming from different sources, e.g., community outsiders.

Several anomaly detection algorithms exist (see Samariya and Thakkar [2021] for a comprehensive review). The method leveraged Isolation Forest due to its simplicity, high performance and because it does not approach anomaly detection solely as the identification of outliers (extreme values) in univariate distributions (Liu, et al., 2018). The broad steps in the algorithm are the following,

1. Pick a random feature (topic) and random value (interaction count) between the minimum and maximum for that feature, using these values to split the dataset.
2. Repeat step (1) for every new split until no further splits are possible (i.e., all data points have been isolated) or if the expanded split trees have some property set according to some heuristic.
3. All data points get a score from the depth of the tree in which they were isolated. The critical insight of the Isolation Forest method is that “normal” data points get isolated in deep trees, while anomalies become isolated in short branches of their own.
4. The method computes an anomaly score for each data point based on this topological property of the split trees. Furthermore, data points with negative scores are labelled as “anomalies” and, conversely, as “normal”.
5. The method changed the score polarity (multiplying by minus one) to make anomaly scores have positive values in the OA typology.
6. The method used the scikit-learn Python implementation of Isolation Forest.
7. The parameter \( \text{random state} = 0 \) was set for reproducibility and \( \text{contamination} = 0.3 \), meaning that up to 30 percent of the time intervals were estimated to be anomalous.
8. The method compared the result with that obtained from setting the contamination parameter to auto, where the algorithm applies its estimation (Liu, et al., 2018).

After the Isolation Forest algorithm scored and classified the time intervals in the dataset, the method applied a second “model explainability” step. The goal was to identify the specific topic interaction patterns that contributed the most to the classification of anomalous time intervals. To inspect the Isolation Forest output model, the method computed the Shapley values from cooperative game theory (Merrick and Taly, 2020). Each topic-interaction count in an interval has a Shapley value, interpreted as a contribution weight toward anomaly or normality.
Results

General description of the topic model

Figure 1 depicts the chosen topic model for the r/digitalnomad corpus (see Methods for details). Panel A, shows the elementary-topic candidates inferred from the corpus vocabulary (depicted by the dots under the ‘Words’ label in the figure). The set of words that belong to some elementary topic appear next to each other and have the same colour. The model’s elementary topic hierarchy contains 83 topics at the bottom level (light blue squares), 16 on the second level (dark blue squares), and two on the third level (red squares), linked to a root elementary topic containing all the words in the corpus vocabulary. For example, the elementary topic (8) in Figure 1A is located in the second level of the hierarchy and it is about interacting with locals in a DN destination country. The words in this elementary topic comprise two percent of the corpus vocabulary. Similarly, Panel B shows the hierarchy of conversation topics. We found 63 such topics at the first level of the hierarchy (light blue squares), 15 at the second level (dark blue squares) and three at the third (red squares). We based the analyses presented in this paper on the second level of the topic model hierarchy (Panel B, dark blue squares).

Figure 1: Topic model for r/digitalnomad corpus. The input documents are the [OC] (title and selftext) shared between 1 January 2019 and 31 December 2021. The topic model has two components (sections labelled A and B). (A) Hierarchical clustering of the corpus vocabulary into elementary-topic candidates. There are 83 candidate elementary topics at the first level (light blue squares), 16 at the second level (dark blue squares) and two at the third (red squares). The most useful keywords to define each elementary topic are in bold. (B) The model's second component groups the threads in the input corpus into (hierarchical) candidate conversation topics. We found 63 topics at the lowest level of the hierarchy (light blue squares), 15 at the second level (dark blue squares) and three at the third (red squares). The raw outputs from each model component were inspected to identify meaningful and coherent elementary topics and topic clusters (see
We kept 14 elementary topics and ten conversation topics after inspection of the model’s output at the chosen hierarchy level. In addition, we highlighted the most informative words to define each elementary topic in Figure 1A. The selected conversation topics have short descriptive titles in Figure 1B. Recall that a conversation topic is a cluster of threads with similar elementary-topic distributions. Different colours in Figure 1B represent the diversity of these distributions within each conversation topic.

Then, we computed the average elementary topic distribution for each topic (using the thread sampling procedure described in Methods) to infer the conversation subjects. Recall that the purpose of labelling and describing conversation subjects is only to provide a broad framework to analyse community interaction patterns, and not to perform content analysis (see Introduction). Table 2 describes the 10 conversation topics. We henceforth use the topic short names in Figure 1 (and also in Table 2) to refer to the conversation topics as quantitative interaction variables.

<table>
<thead>
<tr>
<th>Short name</th>
<th>Threads</th>
<th>Corpus proportion</th>
<th>Interactions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locations</td>
<td>LC</td>
<td>1,271</td>
<td>21.1 percent</td>
<td>39,514</td>
</tr>
</tbody>
</table>

Most threads on this topic fall into two categories: people asking for information about living and working from a location often for a defined period (e.g., a month) or people sharing their experiences in a specific place.

This topic concerns various aspects of work. Primarily, working relationships, e.g., with employers or...
<table>
<thead>
<tr>
<th><strong>Work</strong></th>
<th>WK</th>
<th>1,012</th>
<th>16.8%</th>
<th>25,051</th>
</tr>
</thead>
<tbody>
<tr>
<td>control</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>customers, opportunities for remote workers, and finding temporary jobs. Many threads on this topic are about working skills such as computer programming languages, online teaching and Web development jobs.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Restrictions</strong></th>
<th>RT</th>
<th>777</th>
<th>12.9%</th>
<th>17,871</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The main subject concerns regulations and constraints associated with home countries and DN destinations (but more of the latter), focusing primarily on workarounds to deal with bureaucratic hurdles. There are many threads on local government regulations and negotiating with employers to be able to work remotely. In addition, there are threads about registering a business, how to charge clients, insurance, driving and</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
This topic centres on all visa issues, immigration rules, and what specific visas allow and forbid. While the previous topic concerns a wide range of restrictions, this topic is particular to visas. Conversations are not only about visa restrictions; many threads have a resource-sharing orientation, and others are updates on changing visa rules in different places.

Here we found some of the most introspective conversations in the corpus in which DNs speak about their self-actualisation goals, subjective experiences of local destinations and cultures, and self-expression. A different category of conversations on this topic is...
| Resource sharing | RS  | 506 | 8.4 percent | 10,114 |
| Social interactions | SI  | 304 | 5 percent | 12,334 |

about how to start life as a DN. These conversations mix threads from newcomers and advice from experienced DNs.

Most conversations on this topic are about equipment digital resources, and Internet connectivity, often from the perspective of “how to do things” rather than focusing on the resources as entities per se. Resource-sharing is a key defining characteristic of the DN identity.

The conversations in this topic are often about a DN’s personal relationship with a location — typically with locals, but sometimes with people in co-working spaces, governmental offices and others.

This topic contains conversations
<table>
<thead>
<tr>
<th>Community identities under perturbation: COVID-19 and the r/digitalnomad subreddit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Living online</strong></td>
</tr>
<tr>
<td><strong>Physical address</strong></td>
</tr>
<tr>
<td><strong>Travel</strong></td>
</tr>
</tbody>
</table>

**Exploratory analysis of interaction dynamics**

We now turn to the exploratory analysis of interactions across topics over the three years of this study. We arbitrarily chose to group interaction counts into monthly intervals, using the result to constitute the *topics x interactions matrix, C*, depicted in Figure 2A, using a heat map. In addition, we highlighted critical months in the pandemic timeline using dotted lines, effectively defining three broad periods: (a) *pre-pandemic* (January 2019 to February 2020), (b) *lockdown* (March 2020 to April 2021) and (c) *vaccinated* (May to December 2021).
The pre-pandemic period had lower overall activity compared with the other two periods. Indeed, interaction counts were usually below 2,500 in this period — except for a peak in March 2019 (3,576 interactions). Community interactions peaked again in January 2020, probably due, at least in part, to the coronavirus outbreak. However, the community switched back to the usual pre-pandemic interaction pattern in April 2020, right after the announcement of the COVID-19 pandemic.

There were three other waves of increased overall interaction after the onset of the pandemic. The first peaked in July 2020 (3,897 interactions) — driven primarily by increased VI activity. During the second wave, we observed a new record high in January 2021 (5,357 interactions), from an uptrend that began after the announcement of the first successful COVID-19 vaccine trials on 9 November 2020 (Pfizer, 2020). Finally, activity in the subreddit soared to 8,890 interactions during the third wave peaking in August 2021. After this peak, monthly interaction counts stabilised in the 7,000–8,000 range. The increasingly high interaction waves coinciding with specific critical moments in the pandemic timeline suggest that r/digitalnomad showed clear signs of perturbation.

Figure 2: (A) Heat map of monthly interaction counts across topics. (B) Tree map depicting topic interaction ratios in the three identified sub-periods (pre-pandemic, lockdown and vaccinated).

Note: Larger version of Figure 2 available [here](#).
Another variable we considered in the exploratory data analysis was the set of topic-interaction ratios in the different periods. We computed these ratios and binned the results in 5 percent intervals, aggregating all values equal to or greater than 20 percent into a single bin (see Figure 2B).

Community interactions were skewed almost equally towards LC and WK — both topics were in the 20+ percent bin during the pre-pandemic period. The next most discussed topic was VI (11 percent), then BE, SI, RS and RT (between 5 and 9 percent each) and finally PA, ON and TR (up to 4 percent each). Through the lockdown period, LC remained the most discussed topic in r/digitalnomad. However, we detected some interactional shifts, (a) unsurprisingly, the ratio of WK interactions decreased, while (b) VI became a highly active topic in June and July 2020, and (c) both BE and RT activity increased slightly during the lockdown period. During the vaccinated period, interactions were still skewed towards LC. However, BE became the second most discussed topic, overtaking VI.

Finally, we compared the central tendency of interaction counts for each topic between periods. After confirming that most distributions were not Gaussian (Anderson-Darling, 95 percent CI), we used the Mann-Whitney U test, 95 percent CI for comparing medians, and the FDR p-value correction procedure for multiple testing (Benjamini, 2010). The results are shown in Table 3. We found significant differences for RT, VI and PA between the lockdown and pre-pandemic periods. Furthermore, we found significant differences between the vaccinated and lockdown periods for all topics, except VI and RT.

<table>
<thead>
<tr>
<th>Periods compared</th>
<th>Topic</th>
<th>Median difference</th>
<th>FDR corrected p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-pandemic vs. lockdown</td>
<td>RT</td>
<td>330</td>
<td>0</td>
</tr>
<tr>
<td>Pre-pandemic vs. lockdown</td>
<td>VI</td>
<td>232</td>
<td>0.012</td>
</tr>
<tr>
<td>Pre-pandemic vs. lockdown</td>
<td>PA</td>
<td>23</td>
<td>0.001</td>
</tr>
<tr>
<td>Vaccinated vs. lockdown</td>
<td>WK</td>
<td>369</td>
<td>0.001</td>
</tr>
<tr>
<td>Vaccinated vs. lockdown</td>
<td>RT</td>
<td>364</td>
<td>0</td>
</tr>
<tr>
<td>Vaccinated vs. lockdown</td>
<td>BE</td>
<td>669</td>
<td>0</td>
</tr>
<tr>
<td>Vaccinated vs. lockdown</td>
<td>RS</td>
<td>219</td>
<td>0.017</td>
</tr>
</tbody>
</table>
Community identities under perturbation: COVID-19 and the r/digitalnomad subreddit

<table>
<thead>
<tr>
<th>Vaccinated vs. lockdown</th>
<th>ON</th>
<th>152</th>
<th>0.004</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vaccinated vs. lockdown</td>
<td>PA</td>
<td>113</td>
<td>0.032</td>
</tr>
<tr>
<td>Vaccinated vs. lockdown</td>
<td>TR</td>
<td>207</td>
<td>0.021</td>
</tr>
</tbody>
</table>

The r/digitalnomad OA typology

We computed the discrepancy, $D(x_t)$, for all topics, $x_t \in x$, over $t \in t$ (see Figure 3A). Topics with median discrepancy $|D(x_t)| \leq 0.05$ were considered core, and the remaining were considered peripheral (see Methods for details). The resulting core comprised LC, WK, RT and BE, the remaining six topics were classified as peripheral.

Figure 3: (A) Topic interaction discrepancy box plots. (B) The r/digitalnomad OA Typology. Panel (A) shows the (information theoretic) discrepancy, which measures the degree to which interactions in a given month for some topic, $x_t$, were ‘surprising’ in the context of all interaction counts for $x_t$. Topics with median discrepancy, $D(x_t) \approx 0$ (yellow band), were classified as core, and as peripheral otherwise. Panel (B) is a two-dimensional scatterplot where the interaction patterns for each month, $t$, are plotted using as the $(x,y)$ coordinates the interaction anomaly, $A(t)$, and orientation, $O(t)$, respectively. Months are labelled using the same numbers in Figure 2.

Note: Larger version of Figure 3 available [here].
Figure 3B depicts the 36-month intervals as a two-dimensional points on the OA typology. The X axis is the dimension of interaction anomaly scores, \( A(t) \), and the Y axis is the core-periphery orientation, \( O(t) \) (see Methods for details). In the results presented below, we focused on the OA typology, but in some cases, we also leveraged the detailed interaction patterns depicted in Figure 2 to describe the differences between individual months and observed month clusters in the typology.

To enable nuanced comparisons, we highlighted the three months of the coronavirus outbreak in the pre-pandemic period, i.e., months 12, 13 and 14 using a different colour in Figure 3B. Notice that most pre-pandemic months, including two outbreak months (green and yellow), formed two dense clusters with normal interaction patterns and core orientation. One of these clusters has low anomaly scores, and the other has anomaly scores closer to zero. However, the two clusters are not far apart. We thus henceforth consider a single pre-pandemic cluster.

Three months in the pre-pandemic period deviated from this cluster. The first was month 4 (April 2019), which had an unusually high core orientation. The second was month 5 (May 2019), which was anomalous, but retained a core orientation. However, the anomaly was short-lived. Finally, month 12 (December 2019, outbreak) was similar to month 4.

Seven out of the 14 lockdown months had interaction patterns similar to those observed in the pre-pandemic period. Indeed, interaction patterns in months 16, 17, 21, 26, 27 and 28 expressed the pre-pandemic pattern. Two deviations were caused mostly by pronounced core orientation, namely in months 20 and 22. However, notice that months 12, 18, 23, 24 and 25 formed a different cluster, characterised by a lower (close to zero) orientation, with month 18 having a positive anomaly score. Furthermore, month 19 (July 2020) was not only anomalous, but in that month, the community shifted to a peripheral orientation. However, the last three months of the lockdown period expressed the pre-pandemic pattern again.

The eight months of the vaccinated period all had a clear core orientation, but had anomalous interaction patterns. On average, core orientation was higher than the pre-pandemic average. In addition, we observed that all eight months in the vaccinated cluster are more spread apart in both dimensions of the OA typology. This pattern suggests that the anomalies were unlikely due to just an amplification of the pre-pandemic expression — recall that overall interaction activity was 2.5 times higher than in the pre-pandemic and most of the lockdown period. We saw in the previous section that the interaction-topic ratios changed in the vaccinated period. These shifts are likely causal factors of the observed anomalies in this period.

**Explaining anomalous interaction patterns**

As reported above, one month in the pre-pandemic period, two in the lockdown and all eight months in the vaccinated period had anomalous interaction patterns. Since anomaly scores came from a Machine Learning classification algorithm, we computed the Shapley values from the anomaly classification model to measure each specific topic interaction’s contribution to the scoring and subsequent classification of time intervals as anomalies (see Methods). Figure 4 depicts the Shapley values for all anomalous months in the OA typology.
Figure 4: Shapley values for the interaction counts of anomalous months in the OA typology. The visual
representation of Shapley values for the specific (anomalous) months shows the opposing contributions from topic interaction counts towards anomaly (blue) and normality (red) — ordered from highest to lowest, left to right for contributions toward anomaly (blue), and right to left for contributions towards normality (red).

Note: Larger version of Figure 4 available here.

Our key observations were that (a) the isolated anomalous month in the pre-pandemic period was the result of unusually high WK, low RT and low SI activity; (b) during the lockdown period, VI was a critical conversation topic, which was the primary explanation for the anomalies observed in June and July 2020 — along with increased PA activity; (c) There were several topics that, in different combinations, explained the anomalies observed in the vaccinated period. The only common factor was the high number of LC interactions, which significantly contributed to the classification of seven (out of eight) months in the vaccinated period as anomalous. Indeed, through every perturbation in the three years of Reddit data we analysed, conversations about locations remained the main topic in the r/digitalnomad community core.

Still in the context of the anomalies observed in the vaccinated period, we noticed that the deployment of the first round COVID-19 vaccines, and the subsequent re-opening of borders, coincided with an interaction uptrend on what being a DN means. Increased BE activity contributed significantly to anomalous scores in six of the eight months in the vaccinated period. In five of those months — June to September 2021, and December 2021 — either RS, RT, or both topics showed increased interaction counts, which often contributed to anomaly scores. We observed high BE activity again in October 2021, when VI interactions were also unusually high. PA was another topic contributing to anomaly scores in most of the vaccinated period, coupled with high VI in three months (September to November 2021) in a pattern similar to the anomalies seen in the lockdown period. In addition, elevated PA coincided with increased WK activity in the other three months, namely in May, July and December 2021. Finally, we also noted that ON interactions contributed to anomaly scores in five months (May and September to December 2021), often with elevated VI or WK activity.

Finally, recall from Methods that we set the contamination parameter (that estimated the number of anomalies) to 30 percent based on the exploratory analysis of interaction data. Leaving this estimation to the algorithm resulted in only one difference, namely the classification of June 2020 as normal (but with a score very close to zero). Therefore we were confident in setting the parameter to 30 percent and further the resulting anomalies.

**Conclusions**

**There was a resilient baseline community identity**

This paper began by asking *what emerges* from the conversation threads in the r/digitalnomad community. Relying on an analogy from developmental biology, the answer to this question would be similar to identifying the genes that regulate the identity traits of a living cell — in this study, these genes correspond to the 10 topics identified by the chosen topic modelling algorithm, and the interaction patterns correspond to the expression of identity traits. The second question was whether the dynamic expression of these topics — through community interactions — was stable, thus externalising essential aspects of the community identity.

Using the proposed OA typology, we found that, before the onset of the COVID-19 pandemic, interactions in r/digitalnomad (a) were skewed towards core topics (locations, work, restrictions and being a digital nomad) always in a stable “tension” with peripheral conversation topics, and (b) showed no anomalies.
Therefore we concluded that the community was expressing a baseline identity in that period.

Perturbations affected the regular DN identity expression as the community turned to issues concerning visas temporarily during the lockdown period. However, by the time the first round of vaccination was being deployed and borders began to re-open, r/digitalnomad had recovered its regular core orientation and normal interaction patterns — even after two waves of record-breaking overall interaction counts.

**Post lockdown “chaos”**

The observed expression patterns during the vaccinated period suggest that perturbation sources emerged from the new scenarios created around May 2021 — probably even before, when the first successful vaccine trials were announced. Contrary to the temporary interaction shift to visas during the lockdowns, r/digitalnomad did not show signs of returning to the pre-pandemic identity expression pattern. Indeed, it was moving further away from it in December 2021.

Even though the months in the vaccinated period retained core orientation, the constant anomalous activity was not found to be consistent with an amplification of the pre-pandemic expression pattern. First, notice, in Figure 2B, that the overall interaction ratios across topics changed. Second, the vaccinated months in the OA typology formed a sparse cluster, which can only be the result of different topic-interaction patterns causing the observed anomalies.

The topic interactions that contributed the most to the anomalies in the vaccinated period suggest that DNs were probably dealing with, at least, three complex issues that pushed the community into a state of collective “disorder” or “chaos”. First, the updating the resources and understanding the restrictions that define a new DN emerging from the pandemic. Second, the subject of visas came back in some of the anomalous months. A possible explanation for this is that several countries were probably announcing digital nomad visas or passing new visa laws. Third, the contributions to anomalies related to conversations about physical address and work suggest that remote work might have been a trending subject in this period. However, a more detailed content analysis of the ongoing conversations at the time is outside the scope of this paper.

“**Plasticity**” and transformation of the r/digitalnomad community identity

Still relying on the developmental biology analogy, we observed that the interaction ratios on some topics were “plastic” — they were more variable, particularly under perturbation. For example, conversations about visas, and about being a DN, stand out from all other topics in this regard. Conversely, the expression of topics critical in defining the DN identity, namely locations and resource-sharing, were not “plastic” — their interaction ratios remained relatively constant throughout the three years.

Our results suggest that the r/digitalnomad community probably started a major identity transformation process in May 2021, which was still ongoing in December 2021. The most likely source of perturbation triggering this transformation was an influx of people from outside the community, and not so much the changed conditions and the emerging new “normality” (de Almeida, *et al.*, 2022). Many people in the mainstream lost their jobs and others began to reconsider their values and priorities due to the pandemic. It is reasonable to hypothesise that there was an upsurge of interest in several DN values mostly location independence and resource sharing causing many outsiders to join the community. This would explain the increasing interaction waves in the lockdown, and especially, in the vaccinated period. Certainly, interaction data showed that DNs coped with the effects of lockdowns and mobility restrictions, and that they recovered their identity expression pattern. But the data also showed that a completely changed conversational dynamical landscape emerged in the second half of 2021.
Discussion

Our results agree with most of the findings reported by de Almeida, et al. (2021) and de Almeida, et al. (2022), outlined in the Introduction. First, we showed that, with high probability, there was an increasing influx of aspiring new DNs to the r/digitalnomad community. However, this influx did not affect the community’s identity expression during the lockdown period — only the third and largest wave of interactions in the vaccinated period caused a significant disruption in the expression of the community identity.

Furthermore, de Almeida and colleagues reported an early increase in long introspective DN posts, which is also consistent with the plasticity and elevated activity observed in the BE topic. But, since our analysis covered a more extended period, we saw that this increased BE pattern was amplified in the second half of 2021. Furthermore, we hypothesise that a proportion of this increased BE activity came from outsiders asking, e.g., how to become digital nomads or assessing their skills and education to embark on the DN lifestyle. We formulated this hypothesis from observing the other topics that caused anomalies together with BE in the second half of 2021.

Recall from the Introduction that Zhang, et al. (2017) proposed a synchronic typology to compare Reddit communities in terms of features of their collective identity. One of the specific areas for future work identified by the authors was the study of the temporal dynamics shaping community identity. Here we proposed a diachronic typology that attempts to capture significant dynamic changes in the conversation subjects (core and peripheral) along with interaction anomalies.

As stated in the Introduction, we did not make assumptions about pre-existing community identity, relational expectations between community members or unobservable mechanisms inside people’s minds. Instead, we relied exclusively on externalised interactional data — thus, aligning with theoretical frameworks such as the one proposed by Fuhse (2015). Furthermore, we assumed that the social-semiotic process of identity construction is realised through the dynamics of decentralised information-processing social networks (see Mitchell, 2009, 2006).

Approaching community identity as the emergent dynamical outcome of a complex system allowed us to leverage insights — and get inspiration — from other known complex systems, such as biochemical regulation networks and evolutionary biology. Here we used analogies about how differential gene expression determines cell identity and how some phenotypical traits in an organism are fixed (insensitive to perturbation) or plastic (more variable under perturbation).

The method used in this paper can be applied to the study of any community with almost no changes. No assumptions were made about what constitutes a topic or how many topics exist in a corpus of community conversations. However, this method requires a purposely defined corpus pre-processing and constitution method, as well as the inspection and pruning of both elementary and conversation topics, setting the discrepancy threshold value to classify core and peripheral topics and the contamination parameter value for fine-tuned anomaly detection.

The inferential power of the representations obtained using the proposed OA typology depends, fundamentally, on having data for a sufficiently long period. Indeed, in future work, we will study the r/digitalnomad community for a more extended period before the COVID-19 pandemic, including a more fine-grained content dynamics analysis. Considering such data will allow us to better understand the nature and stability of the baseline community identity, and compare it with the changes observed during the lockdown, vaccinated periods and beyond.

When analysing the dynamics of identity expression over the different periods, we noted that interaction patterns across some topics could have become temporarily correlated, signalling the possible existence of inter-topic synchronisations. An analysis of such synchronisations could reveal emergent self-organisation mechanisms, used by the community to keep the stable expression of their identity, adapt to changing conditions or cope with perturbation.
About the authors

Manuel Pita is an assistant professor of complex systems at CICANT, Universidade Lusófona (Portugal).
E-mail: manuel [dot] pita [at] ulusofona [dot] pt

Karine Ehn is a Ph.D. student at CICANT, Universidade Lusófona (Portugal).
E-mail: karine [dot] ehn [at] gmail [dot] com

Thiago dos Santos is a researcher at COPELABS, Universidade Lusófona (Portugal).
E-mail: thiagohenrique [at] gmail [dot] com

Acknowledgements

Karine Ehn acknowledges funding by the Portuguese Fundação para a Ciência e Tecnologia (FCT), Ph.D. grant UI/BD/151501/2021. Manuel Pita acknowledges Margeret Heath for illuminating discussions and useful suggested edits in the drafts of the manuscript.

References


N. Bozzi, 2020. “#digitalnomads,#solotravellers, #remoteworkers: A cultural critique of the traveling
entrepreneur on Instagram,” *Social Media + Society*, volume 6, number 2 (24 June).


Community identities under perturbation: COVID-19 and the r/digitalnomad subreddit


Editorial history

Received 25 July 2022; accepted 17 July 2022.

This paper is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.
Community identities under perturbation: COVID-19 and the r/digitalnomad subreddit