Universiteit Leiden

Master Computer Science

Uncovering community structures in dark web forums using natural language processing and network analysis

Name:	Filip Jatelnicki
Student ID:	s3122670
Date:	29/08/2024
Specialisation:	Data Science
1st supervisor:	Frank Takes
2nd supervisor:	Gijs Wijnholds
3rd supervisor:	Hanjo Boekhout

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

This thesis examined community structures in dark web forums by integrating Natural Language Processing (NLP), Social Network Analysis (SNA), and sentiment analysis. Our research aimed to provide a comprehensive view of these hidden online communities, capturing the thematic landscape of discussions, underlying patterns of user interactions, and emotional tone of communication.

Addressing our first research question on the dark web's topical landscape, we identified 70 distinct communities of interest through BERTopic modelling. These encompassed a wide range of topics including illicit market transactions, drug trade, security concerns, and social interactions. The topic distribution varied significantly, with clear clustering of related themes. Notably, we discovered a substantial "Outliers" category, highlighting the complexity and diversity of discussions that often defy simple categorization.

In exploring the relationship between topical and structural landscapes, we employed the Leiden algorithm to identify 70 communities of interaction. This analysis revealed that while security and moderation topics appeared universally, some communities specialized in niche areas. The integration of text-based and network-based analyses provided insights into the forum's complex dynamics of information flow and user interactions.

Our investigation into central topics and key influencers utilized centrality analysis, revealing influential users who facilitate information flow across multiple central topics. These users were predominantly involved in moderation, financial fraud, drug trade, and marketplace operations. Interestingly, sentiment analysis showed a uniform distribution of communication tone across both influential and non-influential users, suggesting that centrality does not significantly impact the emotional content of posts.

Comparing the characteristics of the global dark web network with its individual communities, we found complex structures at both levels. While individual communities often formed around specific topics, they still reflected broader forum-wide trends. Both global and local networks exhibited sparse connectivity but significant local clustering and small-world characteristics. Notably, influential users often engaged beyond their community's primary focus, mirroring global network roles.

Our sentiment analysis revealed a prevalence of neutral sentiment with a slight positive skew across the forum, suggesting objective communication despite the often illicit nature of discussions. This sentiment uniformity persisted across communities, reflecting the transactional nature of dark web forums.

This study offers insights into the dynamics of a specific dark web market forum, contributing to our understanding of hidden online communities. While these findings may have implications for cybersecurity and law enforcement, we recognize the need for cautious interpretation and further studies across diverse platforms. The methodologies developed here, combining content analysis with network structure examination, may prove valuable for analyzing other complex online ecosystems, though their broader applicability requires further investigation.

Contents

Chapter 1

Introduction

The Dark Web is a hidden section of the internet that can only be accessed through specialized software. It has become a complex and controversial digital space. Its promise of anonymity attracts a diverse user base, including privacy advocates, political dissidents, cybercriminals, and participants in illicit markets. Dark Web forums, especially those linked to marketplaces, play a crucial role in this ecosystem. They serve as hubs for information exchange, community building, and the facilitation of illegal transactions. This unique ecosystem poses significant challenges for researchers and law enforcement agencies. It requires sophisticated analytical approaches to uncover the complexities of these hidden online communities.

Despite growing interest in Dark Web research, there remains a significant gap in our understanding of the complex community structures within these forums. Previous studies have often focused on either content analysis or network structure, but rarely integrated both approaches.

This thesis addresses this gap by integrating Natural Language Processing (NLP), Social Network Analysis (SNA), and sentiment analysis to investigate community structures within Dark Web forums. Specifically, we examine the Evolution forum, a prominent dark web marketplace operational from January 2014 to March 2015, as a case study. Our research aims to uncover the topical landscape of the dark web and explore how it relates to the structural landscape of the forum's communication network. We investigate which topics serve as central hubs among users and identify key influencers within these thematic clusters. Furthermore, we examine how the structural and topical characteristics of the global dark web forum network compare to those of its individual community networks. By incorporating sentiment analysis, we also explore the emotional tone of discussions across different topics and user groups, providing additional insight into the nature of interactions within these hidden online communities.

To address these questions, we employ BERTopic for advanced topic modeling and the Leiden algorithm for community detection in network analysis. We hypothesize that this integrated approach will reveal complex relationships between topic-based communities of interest and interaction-based structural communities, providing insights into the complex nature of Dark Web forum dynamics.

The remainder of this thesis is structured as follows. First, in Chapter 2, we provide an overview of the necessary background knowledge and preliminary concepts related to the Dark Web, NLP, and network analysis. Then, in Chapter 3, we review the existing literature and related work in these domains, highlighting the gaps that this study aims to address. Chapter 4 presents an initial analysis of our dataset, data preprocessing techniques, including basic statistics and visualizations. This initial analysis lays the groundwork for our more advanced methodological approaches. Chapter 5 outlines our comprehensive methodological approach. It is divided into two main sections: one focusing on the identification of topical communities, and another on the detection of structural communities through network analysis. This chapter explains in detail the specific NLP and network analysis methods applied, as well as how these two approaches are integrated to provide a holistic view of community structures within dark web forums. In Chapter 6, we present the experiments conducted and analyze the results obtained, discussing the findings in the context of our research questions. Finally, Chapter 7 concludes with a discussion of implications, limitations, and future research directions.

Chapter 2

Preliminaries

This chapter presents fundamental concepts essential for understanding complex network structures and textual content analysis. We begin with an overview of key principles in graph theory, including various graph types, metrics, and community detection methods. Following this, we introduce the BERTopic framework (Grootendorst, 2022) for topic modelling, which utilizes techniques such as BERT embeddings, UMAP dimensionality reduction (McInnes et al., 2020), and HDBSCAN clustering (McInnes et al., 2017). The chapter concludes by exploring how graph-based and topic modelling approaches can be integrated, setting the stage for more advanced analyses in subsequent chapters.

2.1 Network notations and definitions

Graph theory (Barabási, 2013) provides a powerful framework for representing and analyzing complex systems, allowing us to model the interactions and relationships between entities as a collection of *nodes* and *edges*. Formally, a *graph* G is defined as an ordered pair $G = (V, E)$, where V is a set of vertices (or nodes) and E is a set of edges connecting pairs of vertices. Each *edge* $e \in E$ is represented as an unordered pair of vertices $e = u, v$, where $u, v \in V$. The *degree* of a node v, denoted as $\deg(v)$, is the number of edges incident to it, which can be formally expressed as $\deg(v) = |u \in V : u, v \in E|$. This relates to the node's connectivity within the graph. Within this framework, a *path* is a sequence of edges that connects a sequence of nodes, where formally, a path P from node u to node v in a graph G is a sequence of edges $(e_1, e_2, \ldots, e_k)$ where $e_1 = (u, v_2)$, $e_k = (v_{k-1}, v)$, and $e_i = (v_i, v_{i+1})$ for each $1 < i < k$. The concept of a path leads us to the definition of *distance* $d(u, v)$ between two nodes u and v, which is the length of the shortest path connecting them. In an *unweighted graph*, this distance is simply the number of edges in the shortest path. Conversely, in a *weighted graph*, where each edge e has an associated weight $w(e)$, the distance is the sum of the weights along the shortest path. Finally, a *component* is a maximal-connected subgraph of a graph. More precisely, it is a subset of nodes $C \subseteq V$ that satisfies two conditions. First, every pair of nodes $(u, v) \in C$ is connected by a path. Second, no additional nodes from $G \setminus C$ can be added to C while preserving this connectivity property.

Graphs can be further classified based on their properties. A *directed graph* has edges as ordered pairs of vertices, indicating a specific direction of the relationship, with an edge $e = (u, v)$ where u is the source vertex and v is the target vertex. In contrast, an *undirected graph* has unordered edges, representing symmetric relationships between pairs of vertices. A *weighted graph* assigns a weight $w(e)$ to each edge e, representing the strength or intensity of the connection. In directed graphs, the shortest path respects the direction of edges; hence, paths are traced following the edge direction from source to target. Conversely, in undirected graphs, paths can traverse edges in any direction. Moreover, in weighted graphs, the shortest path is determined by the minimum summed weight of edges, which may not coincide with the path having the fewest edges. This contrasting approach in unweighted graphs focuses purely on edge count.

In this work, we consider a weighted directed graph and four centrality measures to identify influential nodes within the network.

Degree centrality: Represents the immediate connectivity of a node within the network. For a weighted directed graph, we consider both in-degree and out-degree. Denoted by $\deg_{in}(v)$ and $\deg_{out}(v)$, they are calculated as the sum of weights of incoming and outgoing edges respectively for a node v.

Closeness centrality (Bavelas, 1950): Measure reflecting how close a node is to all other reachable nodes. Denoted by $CC(v)$, it is calculated as:

$$CC(v) = \left(\sum_{u \in V} d(v,u) \right)^{-1}$$

where $d(v,u)$ is the sum of the weights of the edges in the shortest directed path from v to u.

Betweenness centrality (Freeman, 1977): Measures the extent to which a node lies on paths between other nodes. Nodes with high betweenness centrality control the flow of information in a network and can influence the entire network significantly. Denoted by $CB(v)$, it is calculated by:

$$CB(v) = \sum_{s,t \neq v} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}}$$

where $\sigma_{s,t}$ is the number of shortest directed paths from s to t, and $\sigma_{s,t}(v)$ is the number of those paths passing through v, considering edge weights.

PageRank (Page et al., 1998): An algorithm originally used by Google Search to rank web pages in their search engine results. It estimates the importance of a node within a graph based on the incoming links from other nodes. Denoted by $PR(v)$, it is defined as:

$$PR(v) = \frac{1-\alpha}{N} + \alpha \sum_{u \in B_v} \frac{PR(u)}{L(u)}$$

where α is the damping factor, $N = |V|$ is the total number of nodes, B_v is the set of nodes linking to v, and $L(u)$ is the number of links from a node u.

These centrality measures help identify key actors within complex networks, providing insights into their roles and influence in shaping network dynamics and information flow.

Another crucial concept in network analysis is the notion of *communities*, which are subgroups of vertices within a graph that are more densely connected to each other than to other vertices. In directed weighted graphs, this implies that nodes within the same community have stronger and more numerous directed connections among themselves compared to nodes outside the community. To quantify the strength of community structure in directed weighted graphs, we use the *modularity* metric, denoted by Q. For directed weighted graphs, the modularity is calculated using the formula:

$$Q = \frac{1}{m} \sum_{i,j} \left(A_{ij} - \frac{k_i^{out} k_j^{in}}{m} \right) \delta(c_i, c_j),$$

where A_{ij} is the weight of the directed edge from node i to node j, k_i^{out} is the total outgoing weight from node i, k_j^{in} is the total incoming weight to node j, m is the sum of all edge weights in the graph, c_i and c_j are the communities of the nodes, and $\delta(c_i, c_j)$ is the Kronecker delta function, which is 1 if $c_i = c_j$ and 0 otherwise. High modularity values (close to 1) indicate strong community structure, suggesting that the directed connections within communities are significantly stronger than what would be expected by chance, given the nodes' in-degrees and out-degrees.

To detect communities in complex networks, the Leiden algorithm (Traag et al., 2019) is a popular choice. This algorithm offers significant advantages over other community detection methods. Specifically, the Leiden algorithm is known for its ability to overcome issues of non-optimal partitions and disconnected communities that can arise in other methods such as the Louvain algorithm (Blondel et al., 2008). It improves solution quality by refining community partitions more effectively, ensuring communities are well-connected. Moreover, the Leiden algorithm is computationally efficient and scales well with large networks.

The algorithm consists of three phases:

1. **Local moving of nodes**: Individual nodes are moved between communities to improve the modularity score, considering the subgraph consisting of the node's neighbors.

2. **Refinement**: The partition is refined by applying the Louvain algorithm on each community separately, allowing for finding subcommunities within the previously identified communities.

3. **Aggregation**: A new graph is built where each node represents a community from the refined partition. The edges in this new graph reflect the aggregated connections between the communities, with their weights corresponding to the total weight of the edges between the respective communities in the original graph. The algorithm then starts a new iteration with this aggregated graph.

These three phases are repeated until no further improvements in modularity can be made. Compared to the Louvain algorithm, the Leiden algorithm guarantees well-connected communities, provides a more stable optimization, and finds better partitions in fewer iterations.

2.2 Topic modelling with BERTopic

Topic modelling is a type of statistical modelling for discovering abstract topics that occur in a collection of documents, known as a *corpus*. BERTopic is a modern approach to topic modelling that uses advanced natural language processing techniques based on the *transformer model*, specifically leveraging BERT *document embeddings*, which are higher-dimensional vectors representing the semantic meaning of documents. BERTopic clusters similar document embeddings into topics while preserving important contextual information, meaning it considers the surrounding words and sentences to capture the semantic meaning and relationships between words. This allows BERTopic to better understand the nuanced themes and concepts discussed in the dark web forums.

In the BERTopic pipeline, a cluster-based variant of TF-IDF called *c-TF-IDF* is employed to identify the most important words in each topic. c-TF-IDF is a modification of the traditional TF-IDF that operates on a cluster level rather than a document level. It is calculated as:

$$\text{c-TF-IDF}(t, c) = \text{TF}(t, c) \times \log \frac{C}{CF(t)} \tag{2.1}$$

where t is a term, c is a specific cluster, $\text{TF}(t, c)$ is the term frequency of t in topic c, C is the total number of topics, and $CF(t)$ is the number of clusters containing a t. This approach allows c-TF-IDF to identify words that are particularly important within a cluster, relative to their importance across all topics, thus providing a more nuanced representation of each topic's distinctive vocabulary. The notation used in this section differs from that in the Section 2.1 section. Readers should be aware of these differences when interpreting the formulas and symbols presented here.

BERTopic then applies *dimensionality reduction* techniques such as UMAP (McInnes et al., 2020), which reduce the dimensions of embeddings to enhance clustering, followed by *clustering algorithms* like HDBSCAN (McInnes et al., 2017) to form distinctly interpretable topics. Each topic is represented by a set of keywords, known as *topic representation*, that capture the cluster's central theme and summarize the thematic essence of the text clustered together. The quality and interpretability of the generated topics are assessed using *coherence scores*. The entire pipeline of BERTopic is presented in Figure 2.1. By applying BERTopic to analyze the content of dark web forums, we aim to uncover the hidden community structures and interaction patterns that enable illicit activities to thrive in these underground networks. In the next sections, the process of UMAP dimensionality reduction and HDBSCAN clustering will be explained in more detail.

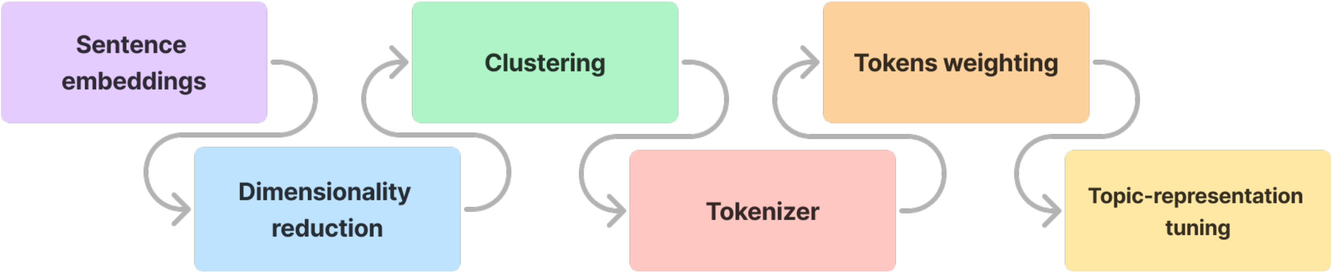


Figure 2.1: General BERTopic approach to topic modelling

2.2.1 Sentence embeddings using BERT

In this section, we delve into the process of creating *sentence embeddings* using BERT within the BERTopic framework. The general overview is presented in Figure 2.2. Sentence embeddings are vector representations of sentences that capture their semantic meaning. These embeddings are essential for understanding the context and relationships among sentences in the text. BERTopic leverages the BERT model to generate these embeddings because BERT, which stands for Bidirectional Encoder Representations from Transformers, excels at capturing the contextual nuances of language through its transformer-based architecture.

The process of creating sentence embeddings using BERT involves several steps:

1. **Tokenization**: The input text is first tokenized using the BERT tokenizer. This process involves splitting the text into individual tokens (words, subwords, or characters) depending on the specific BERT model used. Additionally, special tokens like `[CLS]` (classification token) and `[SEP]` (separator token) are added to mark the beginnings and ends of sentences.

2. **Token embeddings**: Each token is then transformed into a dense vector representation known as a token embedding. These embeddings are learned from BERT's training on large corpora, capturing syntactic and semantic relationships between tokens.

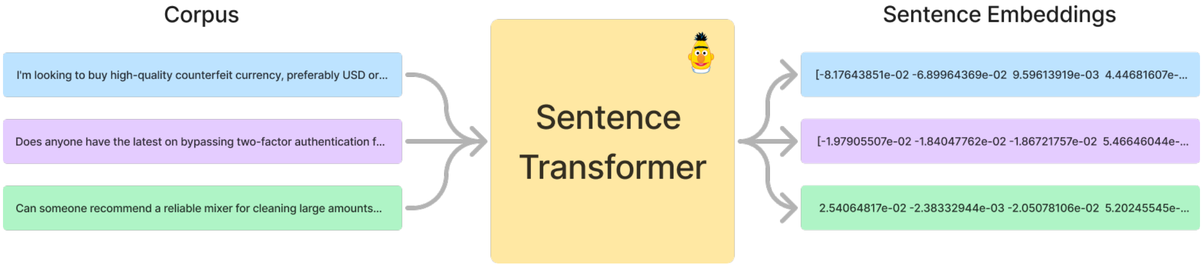

Corpus
I'm looking to buy high-quality counterfeit currency, preferably USD or...
Does anyone have the latest on bypassing two-factor authentication f...
Can someone recommend a reliable mixer for cleaning large amounts...
Sentence Transformer
Sentence Embeddings
[-8.17643851e-02 -6.89964369e-02 9.59613919e-03 4.44681607e-...
[-1.97905507e-02 -1.84047762e-02 -1.86721757e-02 5.46646044e-...
2.54064817e-02 -2.38332944e-03 -2.05078106e-02 5.20245545e-...

Figure 2.2: Illustration of the transformation of text sentences into high-dimensional vector embeddings using a Sentence Transformer. Each sentence from the corpus is processed to produce a unique vector encapsulating its semantic meaning.

3. **Positional embeddings**: To account for the sequence order, BERT incorporates positional embeddings. These learned vectors represent the positions of tokens in the sequence, allowing BERT to understand the context and relational order of tokens.

4. **Input embeddings**: The token embeddings are then combined with positional embeddings through element-wise summation to form input embeddings. These embeddings encapsulate both the meaning of tokens and their positional context within the sequence.

5. **BERT encoder**: Once we have the input embeddings, they are fed into the BERT encoder. The BERT encoder is made up of multiple layers called transformer layers. Each of these layers has a special feature known as *self-attention*.

   Self-attention allows each token in a sentence to look at, or "attend to", all the other tokens in the sequence. For example, in the sentence "The cat sat on the ma", the token for "cat" can attend to "sat", "on", "the", and "mat". This means that every token can learn from the entire sentence, not just from the tokens that are immediately next to it.

   The self-attention mechanism helps BERT capture the relationships and dependencies between tokens across the whole sentence. This ability to understand context is important for grasping complex language patterns. For instance, in understanding that "bank" could mean a financial institution or the side of a river, the self-attention mechanism helps by looking at the surrounding words to decide the correct meaning.

Each transformer layer uses self-attention to weigh the importance of each token in the sequence and how much attention it should give to every other token. This process is repeated across multiple layers, allowing BERT to build a rich understanding of the sentence.

Because self-attention takes into account the entire sequence, it makes sure that the contribution of every token is thoroughly considered. This comprehensive understanding helps BERT generate a fixed-size representation for the entire sequence, ensuring that the essence of the sentence is captured in the representation.

6. **Sentence embeddings**: To generate a fixed-size representation for the input text, BERTopic typically uses the embedding of the `[CLS]` token from the last hidden layer of the BERT encoder. The `[CLS]` token's embedding is considered a representation of the entire input sequence due to its exposure to all tokens through self-attention. Alternatively, pooling strategies like mean pooling or max pooling can be utilized to create fixed-sized sentence embeddings by aggregating information from all token embeddings.

By using BERT's pre-trained model and its advanced self-attention mechanisms, BERTopic can capture detailed and contextual representations of sentences or documents. These representations are in the form of dense vector embeddings, which are rich in meaning and context. This richness helps in effective topic modelling and clustering in later stages of the BERTopic process.

Moreover, BERT has the advantage when handling text data that isn't perfectly clean. Unlike older topic modelling methods that need very clean text, BERT can work well even with messy text data. This makes BERT a powerful tool for dealing with real-world text, which is often not perfectly polished.

The process of creating embeddings may vary in specific implementations, but the core concept remains consistent: BERT generates contextualized token embeddings, which are then aggregated into sentence embeddings. This approach enables rich, context-aware representations of text for various applications.

2.2.2 UMAP

UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2020) is a dimensionality reduction technique used in data science for complex, high-dimensional data. It preserves both global and local structures, making it valuable for visualization, feature reduction, and data preparation for clustering and classification tasks.

Based on principles from manifold learning, UMAP employs advanced mathematical concepts to describe relationships between data points in high-dimensional space. The algorithm starts by identifying neighboring points, then constructs a weighted graph where edges represent similarities between points. UMAP's primary objective is to maintain these relationships as accurately as possible when mapping the data to a lower-dimensional space.

To achieve this, UMAP uses a cost function that minimizes the difference between distances in the high-dimensional and lower-dimensional spaces. This process of creating reduced embeddings is illustrated in Figure 2.3.

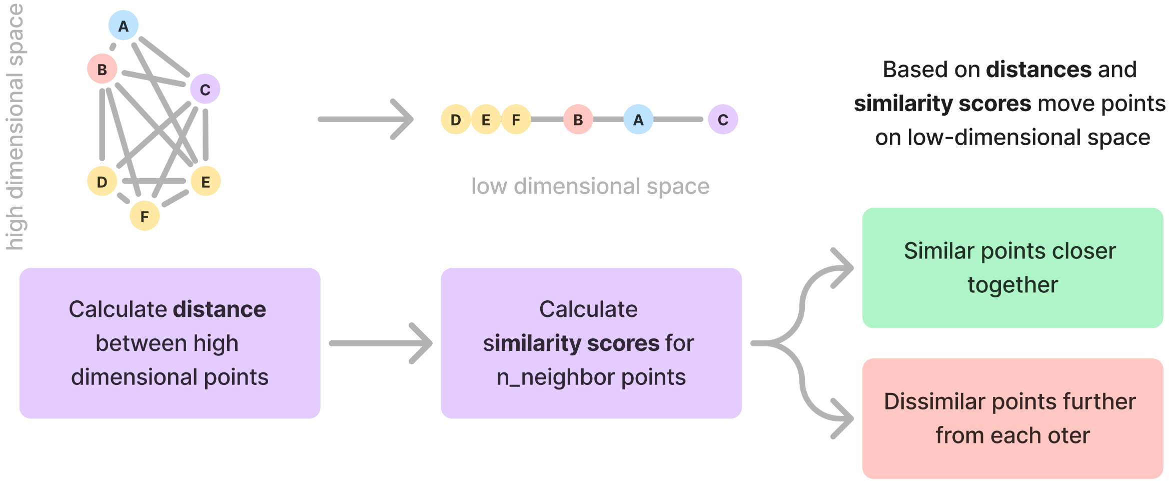

high dimensional space
A
B
C
D
E
F
D E F B A C
low dimensional space
Based on **distances** and **similarity scores** move points on low-dimensional space
Calculate **distance** between high dimensional points
Calculate **similarity scores** for n_neighbor points
Similar points closer together
Dissimilar points further from each oter

Figure 2.3: Illustration of the UMAP algorithm process for dimensionality reduction.

UMAP's efficiency and effectiveness, particularly with large datasets, have led to its widespread adoption in machine learning model feature reduction, data visualization, and exploratory data analysis.

2.2.3 HDBSCAN

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is an advanced clustering algorithm that enhances DBSCAN (Ester et al., 1996) by introducing a hierarchical approach to handle varying densities within datasets. Unlike DBSCAN, which uses a single density threshold, HDBSCAN constructs a hierarchy of clusters across multiple density levels.

The algorithm begins by creating a minimum spanning tree of data points, which is then expanded into a dendrogram representing relationships at different density levels. This hierarchy is condensed into interpretable clusters by identifying the most stable structures across various density thresholds. HDBSCAN automatically determines the optimal number of clusters, eliminating the need for manual specification.

A key feature of HDBSCAN is the "minimum cluster size" parameter, which helps differentiate between noise and significant clusters. This, combined with its ability to evaluate cluster stability across density levels, ensures robust and meaningful results.

HDBSCAN's versatility makes it valuable in various fields, including natural language processing, market research, and social network analysis. Its capacity to handle complex, noisy data with varying cluster densities makes it particularly suited for real-world applications where traditional clustering methods may fall short.

2.2.4 Integration in BERTopic

In the field of natural language processing and network analysis, UMAP and HDBSCAN are often used as part of BERTopic pipeline for efficiently handling and analyzing complex textual data. These techniques play pivotal roles in uncovering underlying structures and patterns within large-scale text corpora:

- **UMAP in BERTopic**: Within the BERTopic framework, UMAP is used specifically to reduce the dimensions of document embeddings generated by models like BERT. This step is crucial for maintaining the semantic relationships between documents while making the data more manageable for subsequent analysis. UMAP's ability to preserve both local and global structures is particularly valuable when dealing with specialized terminology or complex language patterns often found in domain-specific corpora.
- **HDBSCAN in BERTopic**: Following UMAP's dimensionality reduction, HDBSCAN clusters the simplified embeddings into distinct topics. In the context of BERTopic, HDBSCAN's strength lies in its adaptability to the varying densities of topic clusters typically found in natural language data. Its automatic determination of cluster numbers is particularly advantageous when analyzing datasets where the number of topics is not known a priori, such as in exploratory text analysis of diverse corpora.

The synergy of UMAP and HDBSCAN within BERTopic addresses challenges specific to large-scale text analysis, such as handling high-dimensional data, identifying coherent themes in noisy text, and adapting to varying topic distributions. This integration enables researchers to extract meaningful, interpretable topic structures from diverse textual datasets, providing insights into the thematic composition of complex document collections.

2.3 Integrating graph-based and topic modelling approaches

The integration of graph-based and topic modelling approaches provides a comprehensive framework for analyzing complex networks and their associated textual content. This integration combines the strengths of network analysis with the insights gained from natural language processing techniques.

Graph-based approaches allow for the examination of structural relationships within networks, identifying key nodes, communities, and patterns of connectivity. Topic modelling, on the other hand, enables the discovery of thematic structures within large text corpora. By combining these methods, it becomes possible to explore how thematic content relates to network structure, offering a more nuanced understanding of complex systems.

The combination of Natural Language Processing (NLP) and Social Network Analysis (SNA) techniques offers powerful tools for exploring themes, topics, and community structures within networks. For instance, topic modelling techniques like BERTopic can be used alongside community detection algorithms such as the Leiden algorithm to uncover both thematic and structural communities within a network.

Chapter 3

Related Work

The analysis of dark web data is a crucial area of study for researchers, law enforcement, and policymakers. Understanding the dynamics of these hidden online communities is essential for combating cybercrime and other illicit activities that thrive in the anonymity provided by the dark web. Comprehensive analyses of dark marketplaces, such as those revealing significant economic trends and the evolution of practices (Soska and Christin, 2015), underscore the importance of this research. To understand these hidden online communities, researchers have employed techniques from Natural Language Processing (NLP) and Social Network Analysis (SNA). This section explores various approaches to community detection in both traditional social networks and the dark web, highlighting significant contributions and methodologies that inform this thesis on uncovering community structures in dark web forums.

Research in community detection within dark web forums has advanced significantly through classical SNA approaches, beginning with the analysis of the topological properties of dark networks. By examining the structural characteristics of these hidden systems, researchers have gained insights into how they operate (Xu and Chen, 2008). Additionally, detailed case studies of various forums analyze user activity and interaction patterns, outlining the operational structures of these forums (Pete et al., 2020). More practical approaches have further enhanced our understanding. For instance, early warning signals predicting cryptomarket vendor success use SNA to reveal crucial predictive metrics for preemptive action against threats (Boekhout et al., 2024). Furthermore, identifying key players in child exploitation networks aids law enforcement in targeting influential individuals, thereby contributing to practical enforcement efforts (Fonhof et al., 2019). Studies mapping the hierarchy and influence of members using social computing and weighting also contribute significantly to this understanding (Nolker and Zhou, 2005). These studies collectively underscore the value of SNA in comprehending dark web community dynamics and addressing cybercrime.

Integrating NLP techniques into dark web research enriches these analyses. For instance, the analysis of dark websites is critical for developing effective combating strategies against terrorism. A framework to discover latent topics by analyzing the contents of dark websites using Latent Dirichlet Allocation (LDA) (Jelodar et al., 2018) provides insights into the topics discussed by terrorists and extremists (Yang et al., 2009). Recent work has also conceptualized discussions on the dark web using empirical topic modelling approaches, employing various algorithms and coherence measures to uncover thematic patterns in dark web conversations (Basheer and Alkhatib, 2024). Furthermore, sentiment and effect analysis in dark web forums can provide insights into radicalization processes and the emotional states of users, which is critical for identifying emerging threats (Chen, 2008). Additionally, an improved Bi-term Topic Model has been proposed to extract potential topics in darknet market forums, addressing the limitations of traditional topic models like LDA (Yang et al., 2020).

Combining NLP and SNA has led to robust methodologies for analyzing complex data structures within dark web forums. Emotionally driven community detection models that leverage sentiment analysis to identify influential groups within social networks underscore the importance of user behavior and emotional factors in community detection (Kanavos et al., 2017). Enhancements in Twitter community detection incorporate node content and attributes, demonstrating the effectiveness of content-based analysis within SNA frameworks (Alash and Al-Sultany, 2021). Moreover, integrating topic modelling with SNA techniques unveils intricate community structures, further emphasizing the utility of the combined approach (Han et al., 2019).

Our research combines BERT for advanced NLP and the Leiden algorithm for community detection. By integrating these state-of-the-art techniques, our research provides unique insights into the dynamics of dark web communities, possibly enhancing the efficacy of cybersecurity measures and paving the way for more effective interventions against illicit online activities.

Chapter 4

Data preprocessing and initial analysis

In this chapter, we delve into the datasets used for this study, focusing on their origins, structure, and content. We begin by describing the dataset creation process in Section 4.1, detailing the methodologies used for data extraction, structuring, user matching, and network extraction. Section 4.2 provides a comprehensive overview of the text data, including preprocessing steps, analysis of the preprocessed posts, and sentiment analysis. Finally, Section 4.3 then describes the construction and analysis of the communication networks, including the scope of user interactions.

4.1 Dataset origin and information

This research employs a carefully curated dataset from the dark web forum and market Evolution, active from January 2014 to March 2015. This dataset is instrumental in analyzing the complex interactions and behaviors within dark web marketplaces.

The dataset used in this research was obtained from the work of Boekhout et al. (2023), who extracted and structured the data from the Dark Net Market archives (Branwen et al., 2015). The creation process, as described by the authors, involved four key steps:

1. **Data extraction:** Data was extracted from raw HTML files, including forum posts, user profiles, and product listings. The process was organized by date and type of information (forum or market).

2. **Dataset structuring:** After extraction, data was structured into a relational format to resolve issues of duplication, inconsistency, and incompleteness. This structuring was crucial for ensuring data accuracy and usability.

3. **User matching:** Forum user accounts were matched to vendor accounts based on usernames to link forum activities with marketplace behaviors.

4. **Network extraction:** Communication networks were extracted, capturing the interactions among users, to analyze the social dynamics and influence patterns within the forum.

4.2 Text data

As part of our research, we applied a series of preprocessing steps to the raw forum data to ensure data quality, consistency, and privacy protection before analysis. Our preprocessing pipeline, illustrated in Figure 4.1, consists of the following key stages:

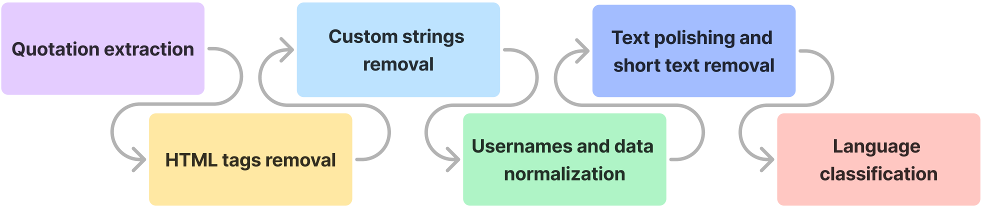


Figure 4.1: Preprocessing pipeline for dark web forum data

Quotes extraction Quotes represent user references or cited material from other users. Posts often include these citations, encapsulated in `<quotebox>`, `<blockquote>`, and `<cite>` HTML tags, presented in Figure 4.2. This makes it easy to extract and remove them from the original content, eliminating redundancy. As mentioned earlier, citations influence the strength of the connection between users while constructing the graph data.

```
<div class="quotebox">
    <cite>QUOTED_USER wrote:</cite>
    <blockquote>
        <div>
            <p>QUOTE CONTENT</p>
        </div>
    </blockquote>
</div>
<p>ORIGINAL CONTENT</p>
```

Figure 4.2: An example of HTML-structured user citations

HTML tags removal After utilizing HTML tags for quotes extraction, HTML tags do not provide any additional semantic information for further analysis. As suggested in (Grootendorst, 2022) documentation, they were also removed completely from the texts using `BeautifulSoup` (Richardson, 2024) library.

Custom strings removal Irrelevant posts and spam content were identified and removed using custom strings collected through regular expressions and other logical methods. Examples of such text are ASCII-art or code snippets posted by users. This step helped to focus the analysis on meaningful user-generated content.

Username and data normalization In posts, usernames were replaced with the placeholder "USERNAME" to protect user privacy and eliminate unusual words. This process involved extracting usernames from the dataset, removing common words like "orange" "blackbird" or "captain" adding absent usernames or users' nicknames, and storing them in JSON format. Usernames consisting of common words were checked against an English dictionary, and any word found in the dictionary was removed from the set of usernames.

Slang terms and various names of drugs and illegal substances were replaced with common words using a custom dictionary created and updated throughout the project. The slang dictionary used for normalizing drug and substance names was compiled using a combination of personal knowledge of slang terms, information gathered from various internet sources, and research conducted directly on the dark web forums being analysed (IACA, 2024), (DarkOwl, 2024). This ensured that the dictionary was comprehensive and reflective of the specific language used within these communities. Slang words were only added to the dictionary when there was a high degree of certainty, erring on the side of caution. In cases of ambiguity, such as the username "MMC" which is also a slang word for Mephedrone, additional rules were implemented to address these issues. This normalization process helped to standardize the vocabulary and improve the coherence of the analysed text without resulting in the loss of context.

Moreover, specific text patterns were replaced with generic placeholders to maintain data consistency. For example, "100 USD" and "100 BITCOIN" were replaced with "PRICE USD" and "PRICE BITCOIN", respectively. Other standardizations included replacing units of measurement with "WEIGHT" or "LITRAGE" dates with "DATE" timestamps with "TIMESTAMP" and edition information with "EDITION". Table 4.1 shows examples of text replacements performed during preprocessing.

Implementing this process using basic Python, concurrent processing or even SQL-based solutions was extremely time and resource-consuming, and after experimentation with various techniques, the most effective solution proved to be the TextSearch library (Van Kooten, 2024).

Original Text	→	**Replaced Text**
100 BITCOIN	→	BITCOIN PRICE
BTC 100	→	BITCOIN PRICE
10 g	→	WEIGHT
3rd ed	→	EDITION
100mg	→	WEIGHT
100$	→	USD PRICE
7321.123	→	NUMBER

Table 4.1: Text replacement examples

Text polishing and short text removal HTML code, PGP signatures, and Bitcoin wallet addresses were removed using regular expressions and custom string matching techniques. Contractions were expanded to their full forms to ensure consistency across the text data.

Texts shorter than 30 characters were removed from the corpus. These short texts often contain noisy reactions, such as "lol" or "yeah", which do not contribute meaningful information to the topic modelling process. By filtering out these short, often irrelevant texts, we focus the analysis on more substantive content, improving the quality and interpretability of the discovered topics.

Language classification Following this step, the text underwent language classification using the fastText model (Joulin et al., 2016), a pre-trained language detection model known for its speed and accuracy. The fastText model was chosen for its ability to handle short texts, support for a wide range of languages, and strong performance on language identification tasks compared to other models (Kostelac, 2021). Only English language texts were considered for further analysis, maintaining the focus on a single language and reducing noise from multilingual content. The distribution of languages found in the dataset is presented in Figure 4.3.

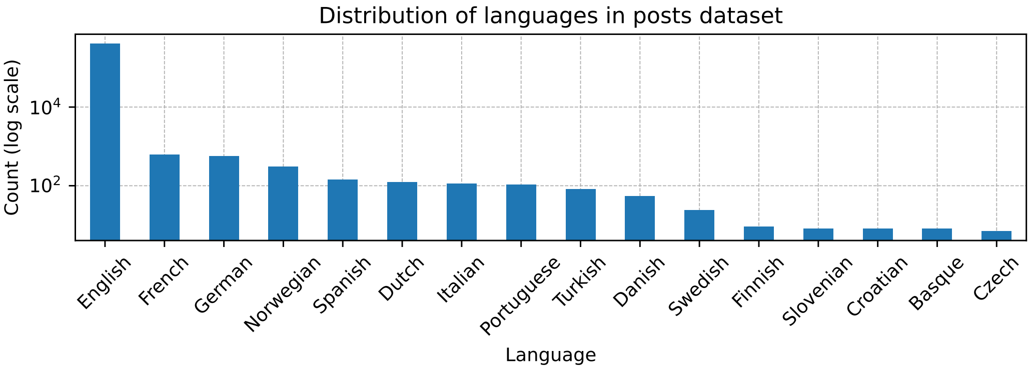


Figure 4.3: Languages distribution in posts dataset

Following this step, the preprocessed data, with cleaned and normalized text, and removed noise, was ready for further analysis using the BERTopic pipeline. Tokenization and feature extraction steps were performed later within the BERTopic framework, as described in the following subsections. By applying these preprocessing steps, the forum data was transformed into a more structured, consistent, and privacy-preserving format, enabling more accurate and reliable analyses of the dark web community's interactions and discussions. The logical flow of the preprocessing pipeline ensured that each stage built upon the previous one, resulting in a comprehensive and efficient data preparation process.

It is important to note that the normalization steps undertaken did not result in the loss of context, as care was taken to ensure that the replacements and standardizations preserved the meaning and intent of the original text.

4.2.1 Preprocessed post analysis

After all preprocessing steps, the dataset consists of 413,267 posts from 20,796 unique users. The text has been cleaned, normalized, and all posts are in English, with slang terms removed. Effects of various preprocessing steps on coherence scores are presented in Figure 4.4.

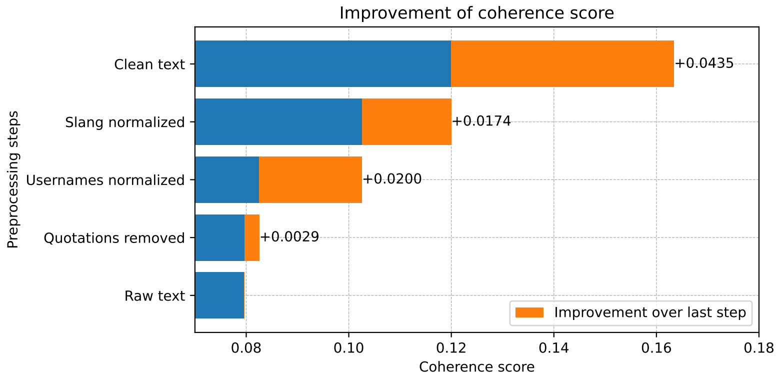


Figure 4.4: Coherence scores at different stages of text preprocessing. The bars represent the coherence scores for raw text, text with quotations removed, text with usernames normalized, text with slang normalized, and fully preprocessed text.

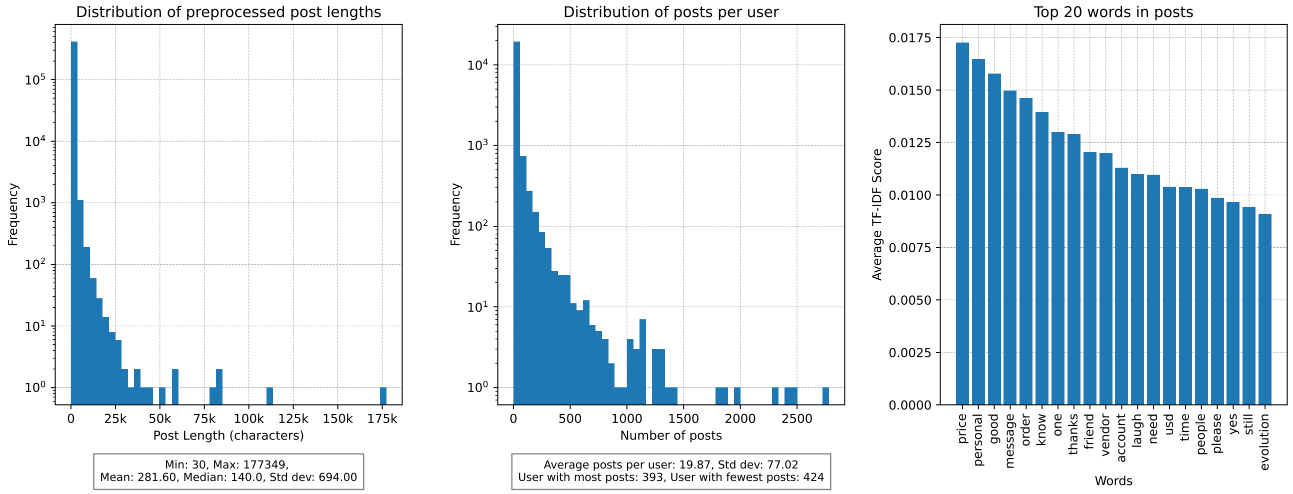


Figure 4.5: Visualizations of post and user characteristics in the dark web forum dataset. (a) Distribution of preprocessed post lengths. (b) Distribution of posts per user. (c) Top words in posts, ranked by TF-IDF scores.

Post length distribution The distribution of preprocessed post lengths, as depicted in Figure 4.5a, follows a heavily right-skewed distribution. This analysis was conducted after removing posts shorter than 30 characters, which is important to note as it affects the overall statistics. Even with this filtering, the distribution indicates that while most remaining posts are relatively short, there are a few posts with significantly higher lengths. The mean post length is approximately 282 characters, with a median significantly lower at 140 characters, highlighting the skew in the data. These values are higher than they would be if sub-30 character posts were included. This suggests that most users tend to write concise posts, possibly for speed or secrecy, which is typical in anonymous online communication platforms. The removal of very short posts allows us to focus on more substantive content while still observing this tendency towards brevity.

Posts per user distribution Figure 4.5b shows the distribution of posts per user, which also follows a right-skewed distribution. This suggests that while the majority of users are less active, contributing fewer posts, a few users are extremely active. These active users, with posts ranging into the thousands, could potentially play central roles within the community, serving as hubs in the network structure of the forums.

Word frequency analysis Figure 4.5c represents the frequency of the top words used in the posts, measured by their Term Frequency-Inverse Document Frequency (TF-IDF) scores. Common words like "order", "price", and "good" have the highest scores, indicating their prevalence across the posts. The frequent usage of such generic terms makes it challenging to draw specific conclusions, except for the observation that the discussions are highly related to orders and transactions.

Additionally, the word cloud illustrated in Figure 4.6 provides a visual representation of the most frequent terms used in discussions within the studied dark web forum. The size of each word indicates its relative frequency, highlighting the most prominent themes and concerns among the participants. Notable terms such as "USE", "PRICE", and "USD" suggest a focus on the commercial aspects of the forum, potentially relating to transactions. Other prominent terms like "message", "account", and "site" reflect common topics related to user interaction and website navigation. This visualization aids in understanding the linguistic landscape of the forum, which is crucial for further natural language processing and social network analysis aimed at uncovering underlying community structures and behaviors.

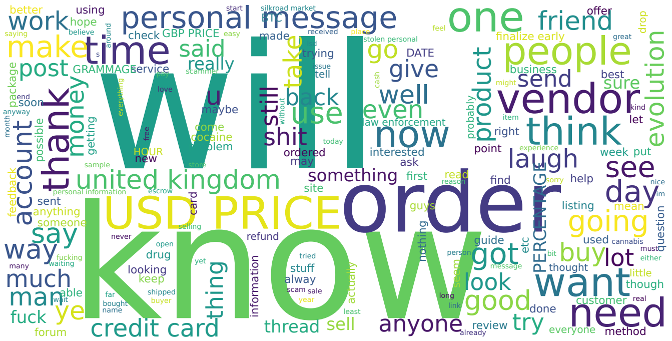


Figure 4.6: Word cloud visualization of the most frequent terms in a dark web forum.

4.2.2 Polarity and subjectivity analysis

Sentiment analysis plays an important role in this study, providing valuable insights into the emotional landscape of dark web forum communications. By employing the the TextBlob library (Loria, 2024), we quantify both the polarity and subjectivity of posts, allowing us to capture the nuanced emotional tone underlying user interactions. This library was chosen primarily for its ability to return both polarity and subjectivity scores. These scores align well with the sentiment distributions we aimed to analyse in this research.

The polarity score ranges from -1 (most negative) to 1 (most positive), while the subjectivity score ranges from 0 (objective) to 1 (subjective). Although more advanced deep learning models may offer higher accuracy, TextBlob provides a reliable and accessible tool for exploring the emotional landscape of the forum discussions.

The sentiment analysis of the forum posts is visualized in Figure 4.7. The polarity distribution (Figure 4.7a) exhibits a sharp peak at zero, suggesting a prevailing neutral sentiment. This finding aligns with the transactional nature of many dark web forums, where participants may intentionally maintain a factual tone to avoid attracting unwanted attention or jeopardizing their anonymity. The distribution shows a slight positive skew, with more posts having positive polarity than negative polarity. There are fewer posts with extreme polarity values (close to -1 or 1), indicating that extreme sentiments are rare. The presence of smaller peaks in the positive and negative regions indicates that a spectrum of emotions is still expressed, albeit less frequently.

The subjectivity distribution (Figure 4.7b) reveals a more nuanced picture. The subjectivity distribution shows a high frequency of posts with a subjectivity score close to 0, indicating that many posts are objective, describing technical details or transaction specifics. The distribution is relatively spread out across the subjectivity range, with posts exhibiting a wide variety of subjectivity levels. There is a broad distribution across the middle range of subjectivity scores (0.3 to 0.7), showing that many posts blend objective and subjective content fairly evenly. This suggests a diverse mix of posts in terms of opinion and factual information. There are fewer posts with very high subjectivity (close to 1), indicating that highly subjective posts are less common.

By applying network analysis techniques to these sentiment patterns, we aim to uncover how the emotional and subjective aspects of the discussions shape the community structures and dynamics within the dark web forums.

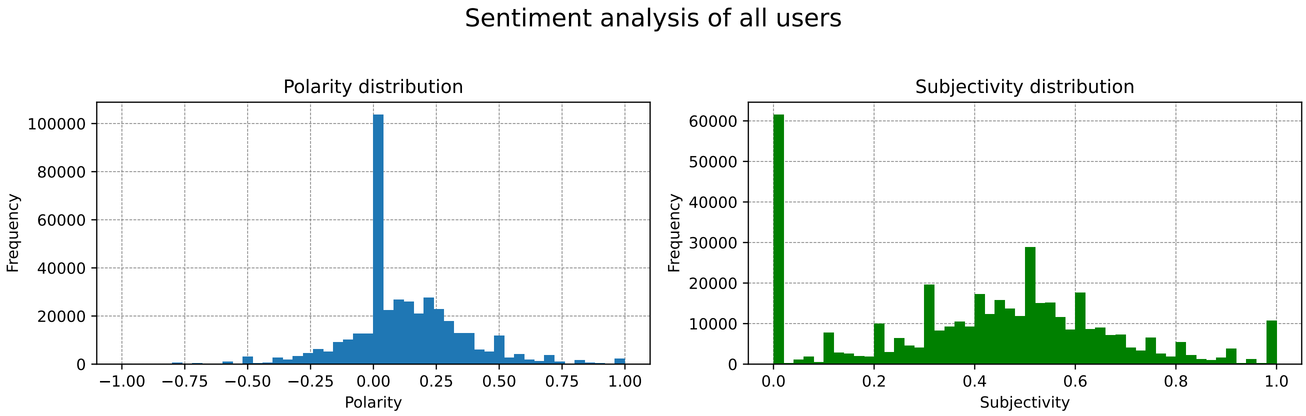


Figure 4.7: Histograms representing the sentiment analysis of forum posts. Figure (a), shows the distribution of sentiment polarity from -1 (most negative) to 1 (most positive), with a notable concentration at 0, indicating a prevalence of neutral sentiments. Figure (b), illustrates the sentiment subjectivity, ranging from 0 (objective) to 1 (subjective), displaying a multimodal distribution with peaks suggesting varying degrees of subjectivity in the posts.

4.3 Graph data

As described in Boekhout et al. (2023), the graph data was constructed from the structured forum dataset to capture the dynamic interactions among users within the Evolution forum. This temporal weighted communication network aims to reflect both direct and indirect interactions among users, as well as shared interests based on their posting behavior in forum topics.

4.3.1 Network construction criteria

The criteria for establishing edges between nodes (users) were based on several parameters:

- **Temporal proximity and sequence**: An edge was created between two users if one user posted after another within a defined time window and sequence order. Specifically, an edge from user a to user b was established if user a posted after user b's post within the same topic, ensuring that user a's post followed user b's post in the temporal order of posts.

- **Weighting of edges**: Edges were weighted based on the time elapsed between connected posts. The weight decreased exponentially with increased time between posts, emphasizing quicker responses, which likely indicate stronger connections or more direct interactions.

The network's edges were assigned weights using an exponential function that considered both the sequence of posts and the time elapsed between them. The formula used to calculate the weight (ω) of an edge from post k by user a to post j by user b is as follows:

$$\omega_{k,a,i,b,j} = \omega_{\text{lower}} + (1 - \omega_{\text{lower}}) \cdot \left(\frac{\exp\left(3 \cdot \frac{\tau_{\text{lim}} + \tau_{k,a,j} - \tau_{k,b,i}}{\tau_{\text{lim}}}\right) - 1}{\mathrm{e}^3 - 1} \right)$$

where $\omega_{\text{lower}} = 0.2$ is the minimum edge weight, $\tau_{\text{lim}} = 7$ days is the time limit for interactions to be considered, and $\tau_{k,a,i}$ and $\tau_{k,b,j}$ are the timestamps of the posts by users a and b, respectively.

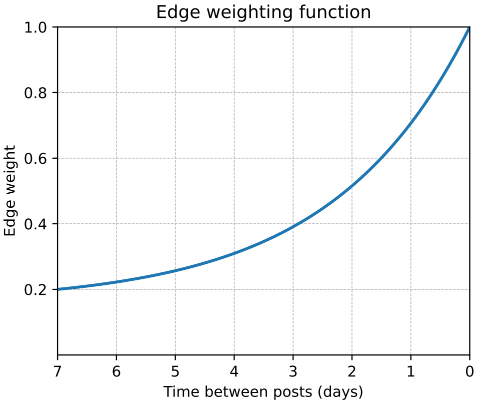


Figure 4.8: Graphical representation of the exponential edge weighting function used in the network construction.

4.3.2 Graph data analysis

The graph is constructed using the `igraph` (Csardi and Nepusz, 2006) library as a directed graph, based on the preprocessed posts' dataset. To ensure a connected structure, we extracted the giant component of the graph, which excludes isolated nodes. This approach enhances the validity and interpretability of centrality metrics and community detection results.

In Table 4.2, we compare network metrics for the full graph and the giant component. The differences are minimal, demonstrating that the giant component effectively represents the overall network. The giant component, mirrors the full graph structure. The network shows a relatively high density with an average degree of 93.64, suggesting active interactions. Despite being weakly connected overall, it lacks strong connectivity; nodes or subgroups can reach others but cannot be reached back, indicating hierarchical or asymmetric information flow.

Metric	**Full graph**	**Giant component**	**Difference**
Nodes	20,793	20,770	-23
Edges	967,289	967,273	-16
Density	2.266×10^{-3}	2.272×10^{-3}	6×10^{-6}
Average degree	93.524	93.641	0.116
Average path length	2.894	2.894	0
Clustering coefficient	0.1446	0.1446	0

Table 4.2: Comparison of network metrics between the full graph and its giant component.

Figure 4.9 illustrates the distribution of centrality measures: The betweenness centrality histogram exhibits a skewed distribution with critical bridge nodes. Closeness centrality shows an almost normal distribution with a slight left skew, indicating uniform node distances. The PageRank distribution is highly skewed, reflecting disparity in node importance. Degree centrality also has a skewed distribution, identifying hub nodes with numerous connections. These histograms highlight important structural properties, such as the presence of influential nodes, the efficiency of information propagation, and the formation of hub nodes that drive community interactions.

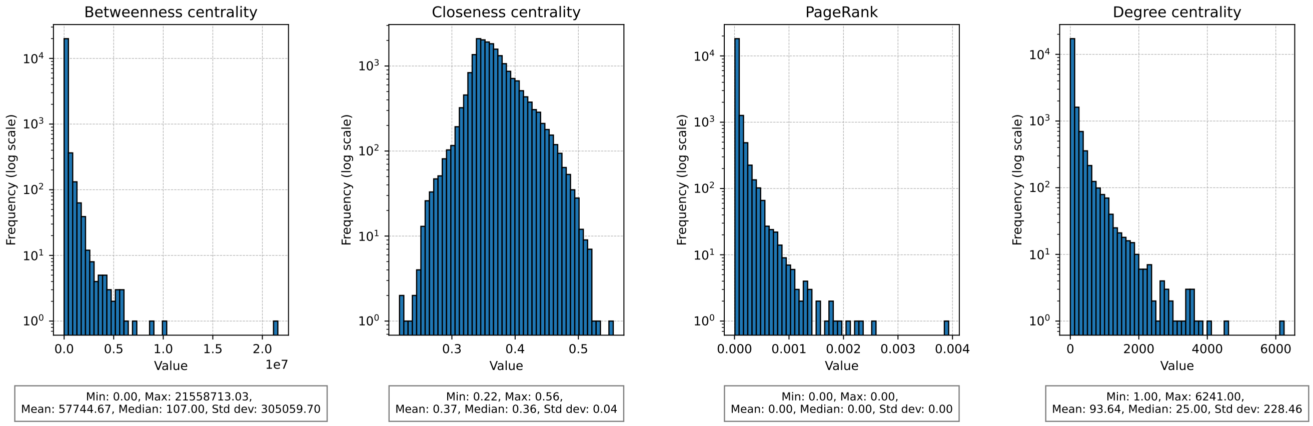


Figure 4.9: Histograms of key centrality measures - betweenness centrality, closeness centrality, PageRank, and degree centrality - for the dark web forum network.

Chapter 5

Methodology

This chapter outlines the comprehensive methodological approach employed in our study of dark web forum communities. This chapter outlines our approach to studying dark web forum communities. We present a detailed account of the techniques, parameter selections, and analytical procedures employed in our research. Our goal is to provide a transparent and reproducible methodology.

In Section 5.1, we discuss section details our approach to identifying communities of interest based on shared discussion topics. We describe the application of state-of-the-art topic modelling techniques, including the use of the BERTopic model, dimensionality reduction, and clustering algorithms. Next in Section 5.2, we explain our methods for uncovering communities of interaction based on user communication patterns. This includes the construction and analysis of communication networks, and the application of community detection algorithms.

The notation used in this section differs from that in the Section 2.1. Readers should be aware of these differences when interpreting the formulas and symbols presented here.

5.1 Topical communities

Our approach to uncovering and analyzing latent topics within the preprocessed dark web forum data utilizes the BERTopic model. This model leverages state-of-the-art language representation and clustering techniques to identify thematic structures in the text. The topic modelling process consists of several key steps, each designed to extract meaningful insights from the complex and often ambiguous language used in dark web forums, as illustrated in Figure 5.1. This pipeline provides a visual overview of the BERTopic process, from input data to final topic representation.

Input data

The input for the BERTopic model is an array of preprocessed English texts. As described in Chapter 4, the raw text data underwent a series of preprocessing steps ensuring that the input texts are clean, standardized, and ready for topic modelling.

Embedding model selection

For our analysis, we employ the `all-MiniLM-L6-v2` model from the HuggingFace Sentence Transformers library (Wolf et al., 2020), the default choice in BERTopic (Grootendorst, 2022). This compact BERT variant offers high-quality sentence embeddings with computational efficiency. Pre-trained on a large English corpus, it maps input text to a 384-dimensional dense vector space.

The model has a maximum input sequence length of 256 tokens. To address this limitation, we implement a text splitting process. Documents exceeding this limit are divided into chunks of fewer than 256 words, split at sentence boundaries to maintain semantic integrity. This approach ensures comprehensive analysis of all content, though it may lead to over-representation of topics from longer posts.

By combining `all-MiniLM-L6-v2` with our text splitting technique, we effectively capture semantic relationships in the dark web forum data. Figure 5.2 illustrates the effect of this process on document length distribution.

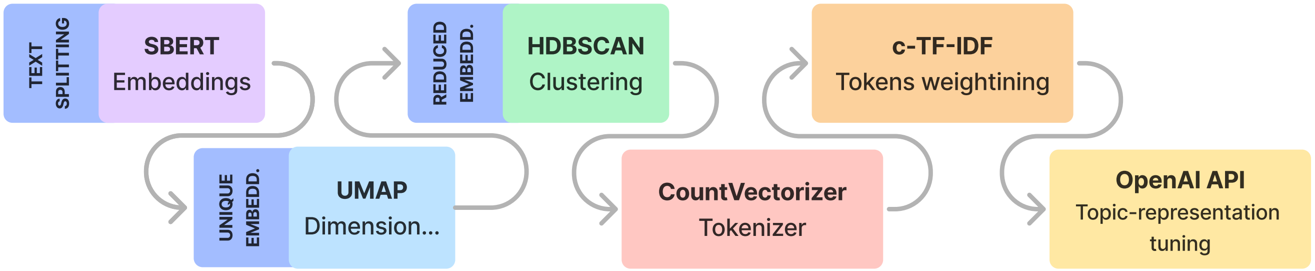


Figure 5.1: BERTopic pipeline

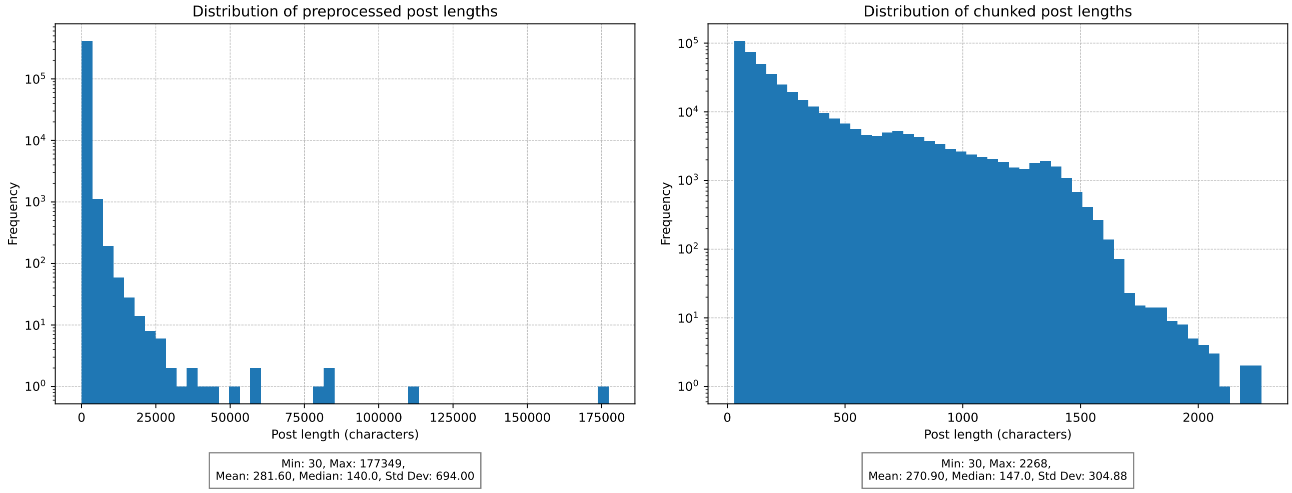


Figure 5.2: Distribution of text lengths before and after chunking. The left plot shows the distribution of preprocessed post lengths. The right plot displays the distribution of post lengths after the chunking process. Both plots use a logarithmic scale.

Unique embeddings

Using unique embeddings or calculating embeddings from unique texts is crucial for improving the quality and efficiency of the topic modelling process (Grootendorst, 2023). This approach allows

Algorithm 1 Split document into chunks

```
procedure SPLIT_DOCUMENT_INTO_CHUNKS($doc, max\_chunk\_length, min\_last\_chunk\_size$)
    $words \leftarrow$ split $doc$ into words    ▷ Split the input document into individual words
    if length of $words \leq 250$ then
        return $[doc]$    ▷ If the document has 250 words or fewer, return it as a single-element list
    end if
    $chunks \leftarrow [,]$    ▷ Initialize an empty list to store the resulting chunks
    for $i \leftarrow 0$ to length of $words$ step $max\_chunk\_length$ do
        $chunk \leftarrow$ join $words[i : i + max\_chunk\_length]$ with space    ▷ Form a chunk by joining words from index $i$ to $i + max\_chunk\_length$
        append $chunk$ to $chunks$    ▷ Add the chunk to the list of chunks
    end for
    if length of last chunk $< min\_last\_chunk\_size$ then    ▷ If the last chunk is too small
        $last\_two\_chunks \leftarrow$ join last two chunks with space    ▷ Combine the last two chunks
        $split\_point \leftarrow$ length of $last\_two\_chunks/2$    ▷ Calculate the split point as half the length of the combined chunks
        replace last two chunks in $chunks$ with:
    • join first half of $last\_two\_chunks$ with space
    • join second half of $last\_two\_chunks$ with space    ▷ Split the combined chunks and update the list of chunks
    end if
    return $chunks$    ▷ Return the list of chunks
end procedure
```

UMAP to more effectively capture the underlying multidimensional structure of the data, resulting in clearer cluster separations and more meaningful visualizations.

While UMAP's author recommends normalizing features (McInnes et al., 2020), our data is already uniformly distributed (Figure 5.3), making normalization unnecessary.

We treated embedding uniqueness as a hyperparameter, systematically evaluating its impact on topic modeling quality. This preprocessing step enhances the interpretability and coherence of discovered topics, improves computational efficiency, and contributes to the robustness of subsequent analyses. Ultimately, this approach leads to more meaningful insights into community structures and interactions within dark web forums.

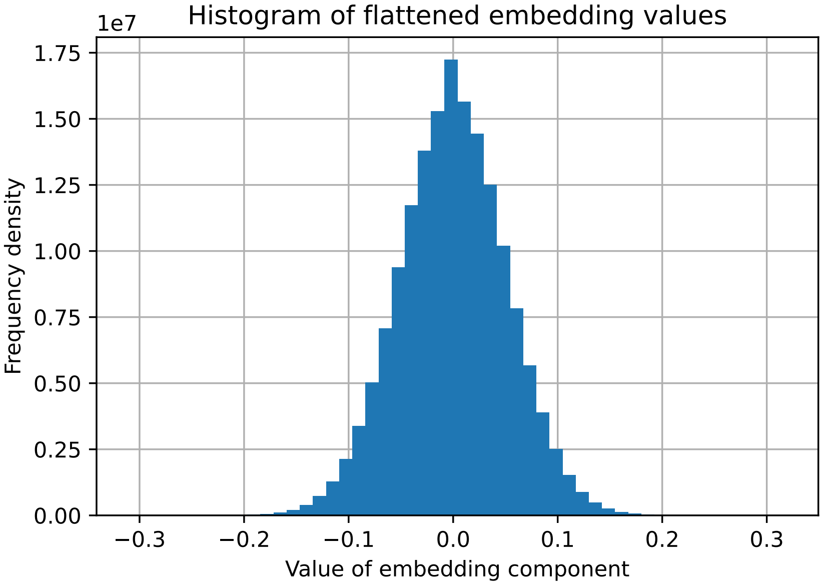


Figure 5.3: Histogram of flattened embedding values: The histogram displays the distribution of embedding component values, showing a bell-shaped curve centered around a mean of approximately 0. The standard deviation of the values is 0.0510, indicating the spread of the data around the mean.

Dimensionality reduction

To visualize and explore the high-dimensional embedding space, we apply UMAP, a non-linear dimensionality reduction algorithm. The embeddings are projected from their original high-dimensional space to a lower dimensional space using the parameters in Table 5.1. The choice of UMAP parameters significantly impacts the quality and interpretability of the low-dimensional embeddings. We carefully select the values based on recommendations from BERTopic and UMAP documentation (McInnes et al., 2020), GitHub issues, and hyperparameter tuning. The search space of hyperparameter tuning is also presented in Table 5.1. Experiments were performed with the RAPIDS CUDA Machine Learning (CUML) (Raschka et al., 2020) implementation of UMAP, which leverages GPU acceleration to significantly reduce computation time.

Setting $n\_neighbors$ to 620 aligns with the heuristic of using $\sqrt{|D|}$, striking a balance between preserving local similarities and revealing overall trends ($\sqrt{|D|} = 642.86$). Projecting to 2D simplifies analysis and visualization while yielding the best results in cluster separation and coherence. The cosine similarity metric captures semantic similarity between text embeddings, and a *spread* of 5 provides clear and visually appealing cluster separation.

The dimensionality reduction results contribute to the thesis goals by enabling visualization of community structures within the dark web forum data. The resulting embeddings support the identification of distinct clusters, representing different communities or topics. This visual representation facilitates understanding relationships and interactions between users and topics.

Furthermore, dimensionality reduction is crucial for efficiently applying the HDBSCAN clustering algorithm. By reducing embedding dimensionality, we make the clustering process more computationally feasible and effective, allowing HDBSCAN to accurately identify dense regions and separate them from noise or outliers, leading to the discovery of meaningful community structures.

Clustering

After reducing the dimensionality of the document embeddings using UMAP, the next step is to identify distinct clusters representing coherent topics within the dark web forum data. For this purpose, we employ the HDBSCAN algorithm (McInnes et al., 2017).

One of the key advantages of HDBSCAN is that it does not require specifying the number of clusters in advance. Instead, it can automatically determine the optimal number of clusters based on the structure of the data. This feature is particularly valuable in the context of topic modeling, as the number of topics is not known beforehand. To find the optimal HDBSCAN configuration for the given data, we conducted a series of experiments using different parameter settings presented in Table 5.1.

We conducted experiments with various HDBSCAN configurations, focusing on the `min_cluster_size` parameter (Table 5.1). Our goal was to balance between capturing general forum topics and avoiding overly specific clusters. We evaluated configurations using the Approximate Density-Based Cluster Validity (ADBCV) score (Halkidi and Vazirgiannis, 2008), which assesses cluster compactness and separation.

Ultimately, setting `min_cluster_size` to 1200 yielded the best results, identifying about 70 distinct, coherent, and interpretable communities with minimal unclassified documents. This configuration effectively captured the key topics discussed in the dark web forum.

Tokenization and tokens weighting

After the HDBSCAN clustering is performed, the next step is to tokenize the documents within each cluster and calculate the importance of each token using the CountVectorizer (Pedregosa et al., 2011) and c-TF-IDF (class-based TF-IDF) algorithms. This process is essential for identifying the most representative terms for each topic, enabling meaningful interpretation of the discovered communities.

The CountVectorizer is used to convert a collection of text documents into a matrix of token counts. It tokenizes the text, learns the vocabulary, and encodes the documents as sparse vectors of token counts. The resulting matrix has one row per document and one column per token (word) in the vocabulary. While the CountVectorizer provides the frequency of each token within the documents, it does not consider the importance of the tokens across different topics. This is where the c-TF-IDF algorithm comes into play. c-TF-IDF is a variant of the traditional TF-IDF weighting scheme, adapted for multi-class scenarios like topic modelling.

As described in the preliminaries (see Chapter 2), c-TF-IDF is a numerical statistic that reflects the importance of a word to a document in a corpus within a specific topic (cluster), considering both the token's frequency within the topic and its rarity across all topics.

By applying the c-TF-IDF weighting, the algorithm identifies tokens that are frequent within a specific topic but rare across other topics. This helps to highlight the most distinctive and representative terms for each community, facilitating the interpretation and labelling of the discovered topics.

The input to the CountVectorizer and c-TF-IDF algorithms is the clustered documents, where each document belongs to a specific topic (cluster) determined by the HDBSCAN algorithm. The output is a weighted matrix of token importance scores for each document within its respective topic. By combining the CountVectorizer and c-TF-IDF algorithms, we obtain a powerful tool for extracting the most salient terms and understanding the key characteristics of each community discovered through the topic modelling process. This information is crucial for interpreting the results, labelling the topics, and gaining insights into the structure and dynamics of the dark web forum discussions.

Topic representation

After extracting the most salient terms for each topic using the CountVectorizer and c-TF-IDF algorithms, the next step is to generate meaningful representations of the discovered topics. These representations should capture the essence of each topic and facilitate interpretation and analysis of the community structures within the dark web forum.

Basic representation

The basic representation of each topic is based on the top keywords extracted using the c-TF-IDF algorithm. The top k terms with the highest c-TF-IDF scores are selected to represent each topic, providing a concise overview of the most distinctive and representative terms associated with each community.

Let $T = t_1, t_2, \ldots, t_n$ be the set of topics discovered by the HDBSCAN clustering algorithm, and let $V_i = v_{i1}, v_{i2}, \ldots, v_{ik}$ be the set of top k terms with the highest c-TF-IDF scores for topic t_i. The basic representation of topic t_i is then defined as:

$$R_{basic}(t_i) = V_i = v_{i1}, v_{i2}, \ldots, v_{ik}$$

This representation allows for a straightforward interpretation of the key themes and concepts discussed within each community.

Maximal Marginal Relevance representation

To enhance the topic representation and generate more informative and diverse keywords, we employ the Maximal Marginal Relevance (MMR) algorithm (Carbonell and Goldstein, 1998). MMR is a technique that aims to select a subset of terms that are both relevant to the topic and diverse from each other, minimizing redundancy in the representation. Given a topic t_i and its associated top keywords V_i, the MMR algorithm iteratively selects terms that maximize a linear combination of relevance and diversity:

$$\text{MMR} = \arg \max_{v_j \in V_i \setminus S} [\lambda \cdot \text{Sim1}(v_j, t_i) - (1 - \lambda) \cdot \max v_k \in S\text{Sim}_2(v_j, v_k)] \tag{5.1}$$

where S is the set of already selected terms, Sim_1 measures the relevance of term v_j to the topic t_i, Sim_2 measures the similarity between terms v_j and v_k, and $\lambda \in [0, 1]$ is a parameter that controls the trade-off between relevance and diversity. By applying the MMR algorithm, we obtain a more comprehensive and diverse representation of each topic, denoted as $R_{MMR}(t_i)$. This representation includes not only the most relevant terms, but also informative keywords that provide additional context and capture the relationships between the key concepts within the topic.

OpenAI GPT API

To generate concise and human-readable labels for each topic, we utilize the OpenAI GPT language model (Radford et al., 2018) through the OpenAI API. The MMR-enhanced topic representations $R_{MMR}(t_i)$ are used as input for a carefully crafted prompt that instructs the GPT model to generate a trio of capitalized words encapsulating the general themes of the texts.

The prompt, as shown in Algorithm 2, leverages the GPT model's ability to understand context and generate semantically meaningful text based on the given input. By providing the MMR-enhanced topic representations as input, the GPT model synthesizes the key information and generates concise topic labels that capture the essence of each community. The OpenAI API allows for seamless integration of the GPT model into the topic representation pipeline. By making API calls with the appropriate prompts and input data, we can efficiently generate topic labels without the need for local model deployment or extensive computational resources.

Let $R_{GPT}(t_i)$ denote the topic label generated by the GPT model through the OpenAI API for topic t_i based on the MMR representation $R_{MMR}(t_i)$. This final representation provides a human-readable overview of the main themes and concepts discussed within each community, facilitating the interpretation and analysis of the discovered topics.

Algorithm 2 OpenAI GPT API call for topic labelling

```
Prompt ← "You are an assistant skilled in summarizing and labelling topics based on keywords
    extracted from documents. Your task is to generate a trio of words that encapsulate the
    general themes of the texts. Format your response as a comma-separated list with each
    keyword capitalized, ensuring they accurately reflect the content. Example: Health, Education,
    Technology"
Input ← $R_{MMR}(t_i)$
$R_{GPT}(t\_i)$ ← OpenAI_API(Prompt, Input)
```

By combining the basic representation, the MMR-enhanced representation, and the GPT-generated topic labels obtained through the OpenAI API, we obtain a comprehensive characterization of the topics discovered within the dark web forum.

Limitations

It is important to acknowledge some limitations of the topic modelling approach used here. The identified topics are ultimately a function of the specific algorithm and parameters employed, and alternative approaches may yield different results. Additionally, while the topics provide a useful overview of the content landscape, they do not in themselves fully capture the nuances and complexities of the forum interactions.

Nonetheless, the topic modelling results offer a valuable starting point for understanding the key dimensions and organizing themes of the dark web forums. They provide a basis for identifying areas of focus for further analysis, and for beginning to map out the relationships between different domains of activity and interaction.

5.2 Structural communities

This section outlines our methodology for identifying and analyzing structural communities within the dark web forum. Our approach focuses on uncovering distinct community structures based on user interactions and communication patterns. A key aspect of our method is the application of graph theory and network analysis techniques to detect communities of interaction.

We employ a multistep process that involves constructing interaction networks, applying community detection algorithms, and optimizing parameters to identify meaningful community structures. Our goal is to reveal the underlying social structure of the forum, complementing the topical analysis described in the previous section.

A particular focus of our methodology is to align the graph-detected community structures with the topical communities identified through text analysis, aiming for a cohesive understanding of both the content and interaction patterns within the forum. The following subsections detail each step of our structural community detection process, including the variants of graph representations considered, the algorithms applied, and the optimization techniques used.

Graph variants To achieve robust community detection results, we analyzed multiple variations of the Giant component of the interaction network:

- Original interaction weights.
- Normalized weights (both directed and undirected).
- Uniform weights to assess the impact of weight variations.

Weights normalization Normalization was used to mitigate the impact of outliers, making weight comparison across the network simpler (see Figure 5.4). Standardizing weights (scaling them between 0 and 1) reduced variability, thus enhancing the comparability and reliability of the statistical analysis.

Partitioning algorithms We applied the Leiden algorithm for community detection, focusing on maximizing modularity, and adjusting the resolution parameter to identify approximately 70 communities, deliberately aligning with the number of topics uncovered through our BERTopic analysis. This alignment allows for a direct comparison between the topical structure and the interaction-based community structure of the forum.

Our process began with a modularity-based approach to identify suitable network variants. Subsequently, we fine-tuned the resolution parameter to achieve the target number of communities. Multiple iterations of the Leiden algorithm were performed to ensure robust results, with careful parameter adjustments allowing for detailed control over the community detection process. This iterative approach helped us maintain consistency with our topical analysis while ensuring the structural integrity of the detected communities.

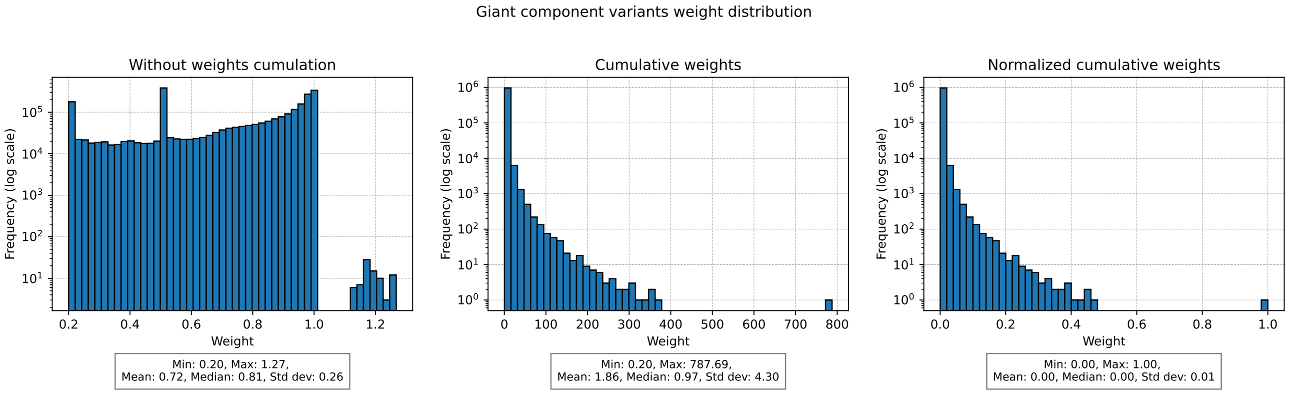


Figure 5.4: Weight distributions: (a) Original, (b) Cumulative, (c) Normalized cumulative weights.

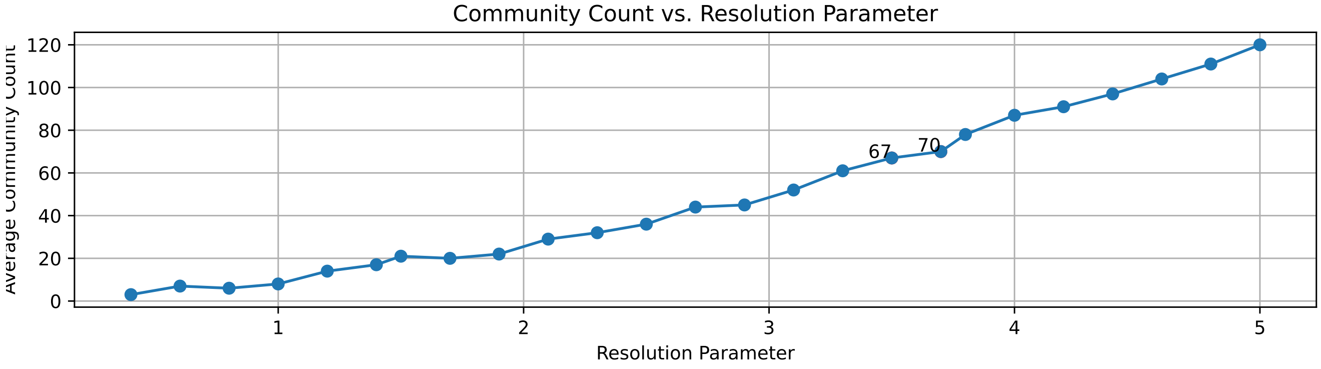


Figure 5.5: Community count vs. resolution parameter.

Graph type	**Direction**	**Weight**	**Modularity**	**Community count**
Full graph	Directed	Original	0.447028	27
Giant component	Directed	Original	0.447528	9
Giant component	Directed	Uniform	0.447360	11
Giant component	Undirected	Original	0.442593	7
Normalized giant component	Directed	Normalized	0.457707	7
Normalized giant component	Undirected	Normalized	0.456376	7

Table 5.2: Graph configurations: modularity scores and community counts

Resolution value optimization The resolution parameter was varied between 0.4 and 5. We found that a resolution value of 3.7 resulted in the target of around 70 communities (see Figure 5.5). However, this configuration yielded a modest modularity score of 0.2313.

The choice of resolution value highlighted the trade-off between community count and modularity. Despite the ability to meet the quantitative target of approximately 70 communities, there was no significant optimization for qualitative community structure.

Modularity scores and community structures were analyzed across graph configurations (Table 5.2). Normalized weight graphs achieved higher modularity scores, highlighting the significance of weight normalization in detecting coherent communities. For example, the directed normalized Giant Component scored 0.457707 in modularity.

In conclusion, weight normalization proved crucial for effective community detection, revealing coherent community divisions within the network. Further analysis and discussion of these communities and their structures are detailed in subsequent chapters.

Parameter	**Values tested**	**Selected value**	**Rationale**
UMAP			
unique_embeddings	True/False	True	Unique embeddings might help in better UMAP results.
text_case	Lowercase, Original		Lowercasing simplifies the model processing. Proved to be beneficial during the language detection phase.
n_components	[2, 3, 5, 7, 10, 17, 20]	2	Number of dimensions for projection; fewer dimensions might lead to loss of information, while more could preserve too much noise.
metric	[cosine, euclidean]	cosine	distance calculation method in high-dimensional space, impacting clustering quality.
spread	[1, 3, 5]	5	Controls how clustered the embedded points are; a higher spread could lead to more diffuse clusters.
n_neighbors	`range(100, 821, 40)`	620	Local versus global structure in the data. Larger values prioritize global structure which can be crucial for identifying less obvious but relevant patterns.
HDBSCAN			
min_samples		1	Minimum number of samples in a neighborhood for a point to be considered a core point
cluster_selection_method		“eom”	Method for selecting clusters from condensed tree
gen_min_span_tree		True	Generate the minimum spanning tree for the data
approx_min_span_tree		False	Approximate minimum spanning tree algorithm
min_cluster_size	`range(100, 1600, 50)`	1200	Minimum data points required to form a cluster

Table 5.1: BERTopic hyperparameters search space and rationale

Chapter 6

Experiments and results

This chapter presents our experimental analysis of community structures within the Evolution dark web forum. Our approach integrates advanced topic modelling techniques with network analysis to provide a comprehensive understanding of both the thematic landscape and the structural relationships within the forum. Through a series of analyses, we explore the topical landscape, examine the interplay between content and network structure, investigate the characteristics of influential users, and compare local community features to the global forum landscape.

The findings presented here offer multifaceted insights into the complex dynamics of dark web forums, revealing both thematic patterns and structural relationships that characterize these hidden online spaces. Each section addresses specific research questions, which are introduced at the beginning of the respective analyses. These results form the basis for our subsequent discussions on the implications for understanding dark web communities and potential interventions in illicit online activities.

6.1 Topical landscape

In this section, we analyze the Evolution dark web forum dataset to uncover the various themes and topics discussed within this hidden online community. Our primary research question is: *What does the dark web's topical landscape look like?* To address this, we employ BERTopic, a state-of-the-art topic modelling technique that combines advanced language representations with clustering algorithms.

Through rigorous preprocessing, embedding, and clustering processes, we systematically uncover and analyze latent topics in the forum posts. This approach identifies key discussion themes and illuminates the intricate community structures and interaction patterns characterizing these forums. Figure 6.1 visualizes the distribution of topics in a reduced two-dimensional space, offering a comprehensive overview of the topical landscape and Figure 6.2 visualizes the distribution of topics across the corpus, emphasizing the significance of each topic.

As seen in Figure 6.1, the topics form distinct clusters, with some closely related topics positioned near each other. For instance, we can observe clusters related to financial transactions (”Bitcoin, Exchanges, LocalBitcoins”, ”PayPal, Transfers, Accounts”), drug-related discussions (”Cannabis, Strains, Smoking”, ”Stimulants, Opioids, Drugs”), and marketplace operations (”Orders, Shipping, Delivery”, ”Feedback, Buyers, Vendors”). This visualization helps us understand the thematic structure of the forum discussions and the relationships between different topics.

Our analysis aims to identify and understand the key topics of discussion, explore relationships between topics, and provide insights into the nature and dynamics of dark web content. We seek to reveal the complex ecosystem of illicit activities, security concerns, and social interactions, ultimately contributing to our understanding of community structures and dynamics in dark web forums.

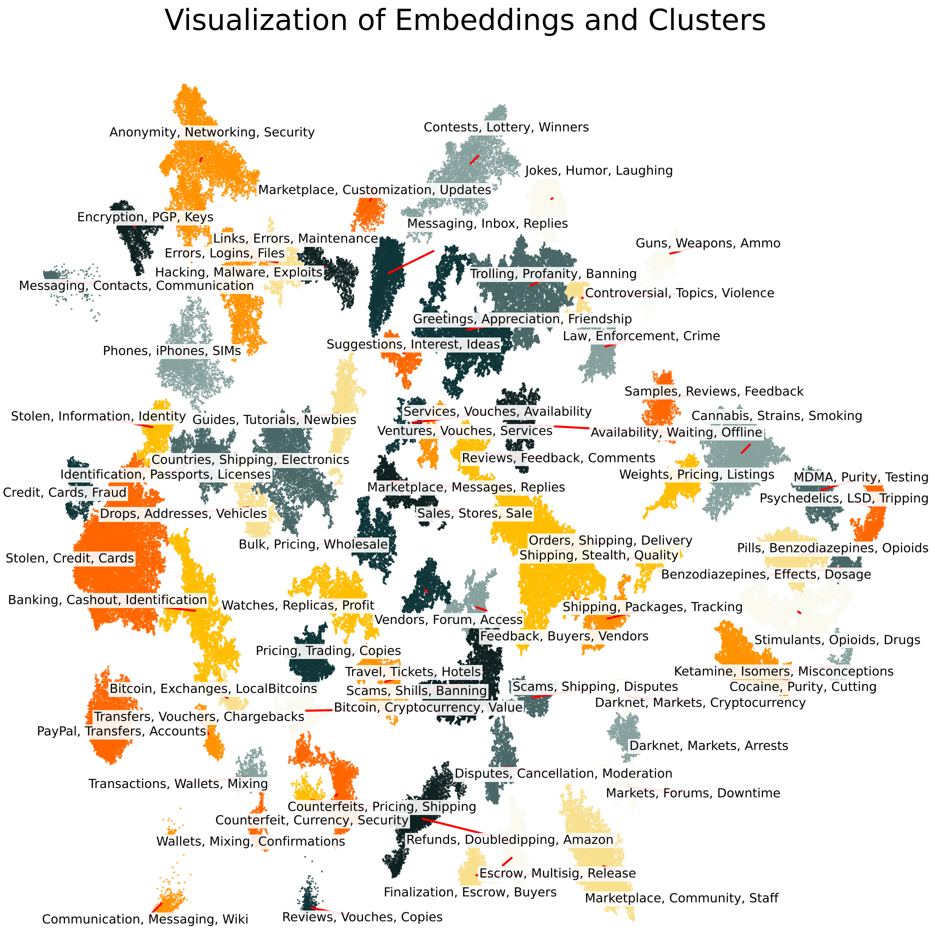


Figure 6.1: Visualization of topic embeddings in reduced 2D space. Each point represents a document, color-coded by its assigned topic. The proximity of points indicates semantic similarity between topics.

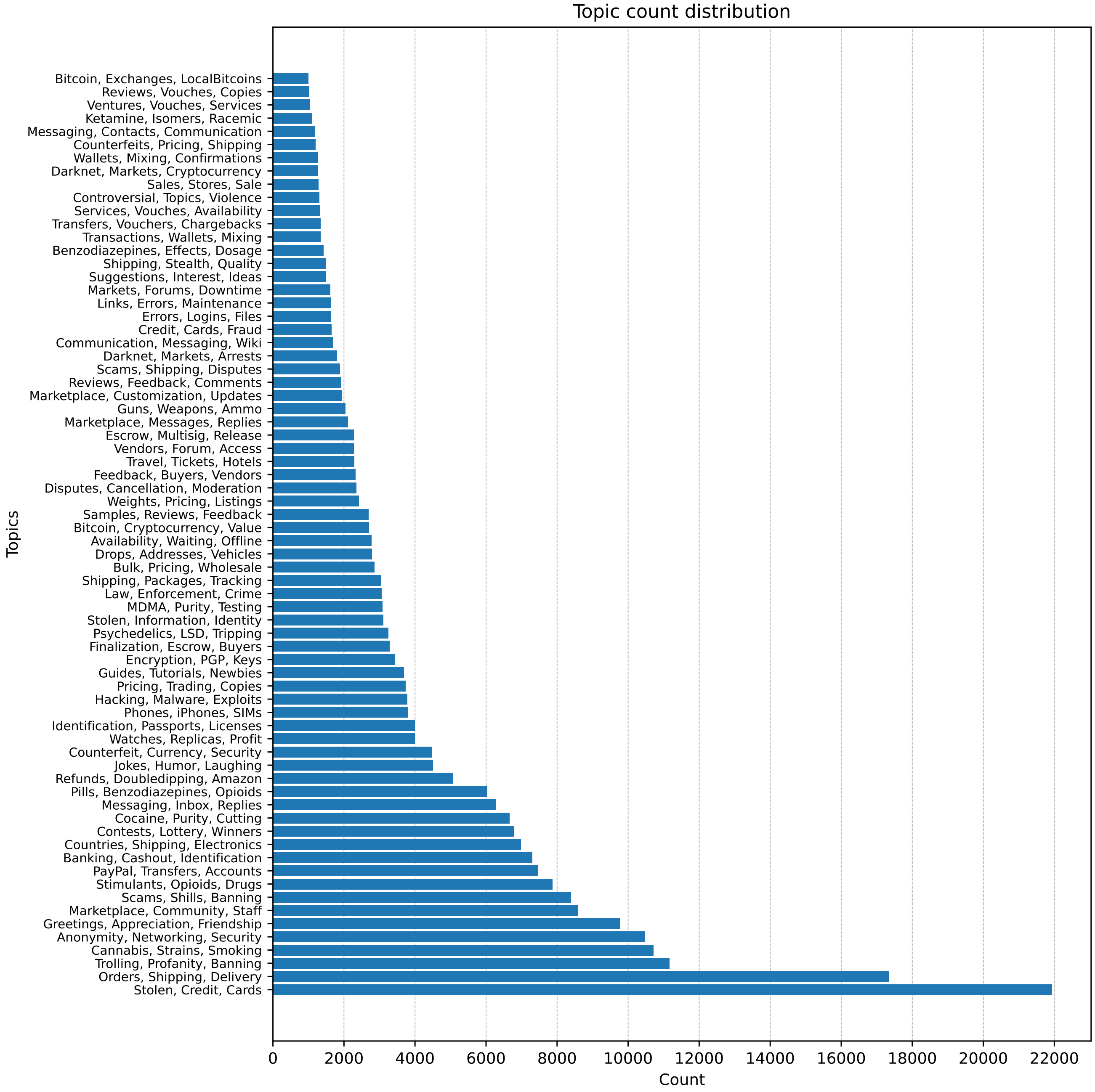


Figure 6.2: Topic distribution

6.1.1 Outliers

The analysis revealed a significant “Outliers” topic comprising 145,808 posts, which accounts for 34.17% of the entire corpus. It is crucial to note that this substantial outlier category has been excluded from Figures 6.1 and 6.2 for clarity of visualization. Despite its absence in these figures, the outlier topic far exceeds any other identified topic in terms of post count, significantly impacting the overall distribution of topics in the dataset. This outlier category appears to capture a considerable amount of content that falls between or outside the more well-defined categories identified in the analysis. While preprocessing steps were taken to remove short, random texts and simple reactions, the prominence of the “Outliers” topic suggests that a large portion of the forum posts may be characterized by more ambiguous, peripheral, or hard-to-categorize content. The existence of such a substantial outlier category highlights the diverse and often ambiguous nature of discussions in dark web forums, as well as the challenges in applying traditional topic modeling techniques to this type of data.

Examining a sample of posts assigned to this topic reveals a mix of content types, including side discussions, vague or incomplete statements, and posts that touch on multiple themes without clearly fitting into any single category. For example, some representative posts include: `‘‘I heard something about that, but I’m not sure’’` or `‘‘Has anyone tried this before? Thoughts’’`. These types of posts, while not entirely deprived of meaning, do not align with the more specific and focused topics identified in the analysis.

It is worth noting that some texts categorized as outliers may be considered as such due to their lack of context. For instance, a post like `‘‘Has anyone tried this before? Thoughts?’’` could potentially be referring to a wide range of subjects, such as drugs, tutorials, services, or other topics. If the necessary contextual information were available, these posts could likely be more accurately classified under the appropriate topic categories. The absence of clear context in many of these outlier posts highlights the challenges inherent in analyzing complex, semi-structured, online discourses, where the full meaning of a given post may depend heavily on its position within a larger conversation or thread.

The high prevalence of these “Outlier” posts suggests that much of the forum activity may be characterized by more open-ended, exploratory, or unfocused discussions, rather than strictly goal-oriented or transactional exchanges. This finding highlights the complex and diverse nature of communication within dark web forums, where participants may engage in a wide range of discourse modes and styles that extend beyond the core topics and activities of interest.

6.1.2 Topic distribution and relationships

Our analysis revealed a diverse range of topics within the Evolution dark web forum. As shown in Figure 6.2, the distribution of topics varies significantly, with some dominating the discourse while others appear less frequently.

Figure 6.1 provides a comprehensive overview of the topical landscape, visualizing the distribution of topics in a reduced two-dimensional space. This visualization helps us understand the thematic structure of the forum discussions and the relationships between different topics.

Transactional aspects of dark web marketplaces are prominent, with topics like “Stolen, Credit, Cards” (count 21,938) and “Orders, Shipping, Delivery” (count 17,345) among the most frequent. Various payment-related topics such as “PayPal, Transfers, Accounts” (count 7,469) and “Bitcoin, Cryptocurrency, Value" (count 2,711) also feature prominently, underscoring the centrality of commerce and financial transactions in these forums.

Drug-related topics form another significant cluster, including "Cannabis, Strains, Smoking" (count 10,717), "Stimulants, Opioids, Drugs" (count 7,872), "Cocaine, Purity, Cutting" (count 6,668), and "Pills, Benzodiazepines, Opioids" (count 6,033). The prevalence of these topics highlights the significant role of drug trading in shaping forum content and interactions.

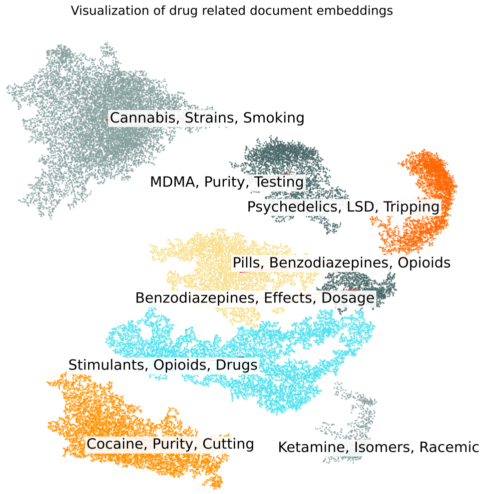


Figure 6.3: Drug related document embeddings visualized on two-dimensional space

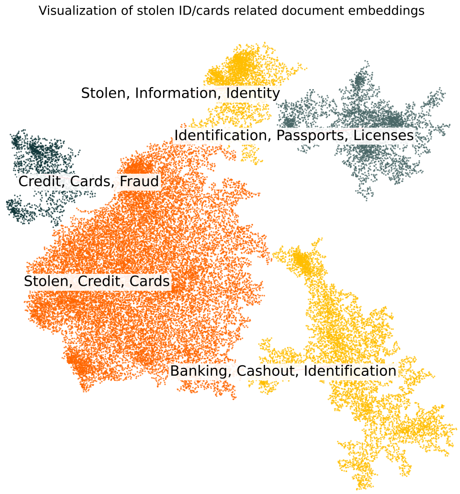


Figure 6.4: Stolen documents related document embeddings visualized on two-dimensional space

Figures 6.3 and 6.4 present zoomed-in portions of the embedding space, focusing on drug-related topics and stolen documents/credit cards, respectively. These visualizations demonstrate clear boundaries between clusters while also revealing the proximity of intuitively related topics. Both figures represent magnified views of specific areas within the complete 2D embedding space depicted in Figure 6.1.

Security and trust-related topics are also prominent, including "Anonymity, Networking, Security" (count 10,462), "Scams, Shills, Banning" (count 8,398), and "Encryption, PGP, Keys" (count 3,439). These topics underscore the crucial role of maintaining security and reducing risks in an environment where many participants are engaged in illicit activities.

Interestingly, social and communal aspects of the forums are represented in topics such as "Greetings, Appreciation, Friendship" (count 9,769) and "Jokes, Humor, Laughing" (count 4,509), indicating that forum members engage in social bonding and identity formation alongside more instrumental activities.

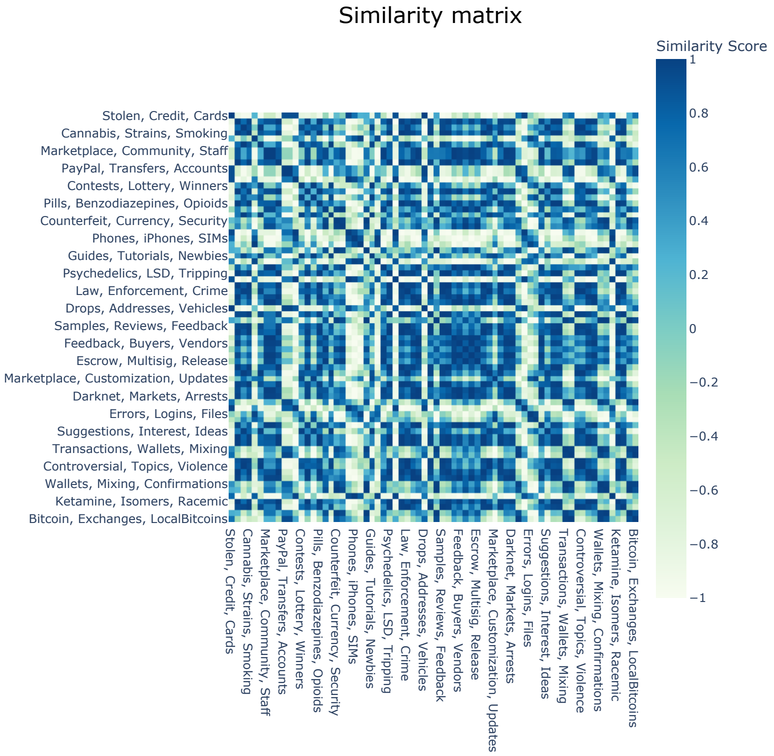


Figure 6.5: Similarity matrix of topic embeddings. Darker shades indicate higher cosine similarity between topics.

Figure 6.5 presents a similarity matrix that displays the pairwise cosine similarity scores between the identified topic embeddings. This visualization helps illustrate the thematic relationships and overlaps between different areas of discussion. The matrix reveals strong similarities between various drug-related topics, as well as between topics related to marketplace mechanics and security.

Figure 6.6 shows the hierarchical clustering of topics, revealing key themes and areas of focus. One prominent cluster centers around the mechanics and dynamics of the dark web marketplaces themselves, with topics like “Marketplace, Community, Staff” (count 8,595), “Samples, Reviews, Feedback” (count 2,698), “Vendors, Forum, Access” (count 2,285), and “Marketplace, Messages, Replies” (count 2,112). These topics point to the importance of the marketplace platforms in facilitating and structuring user interactions.

The relationships between these various topic clusters provide insights into the structure and dynamics of the dark web forum communities. For example, analyzing the co-occurrence patterns of drug-related topics with those related to transaction mechanics and security could shed light on how trust and risk are negotiated in the context of illicit market exchanges. Similarly, examining the interplay between social bonding topics and those focused on commerce and security could help to understand how community norms and relationships shape and are shaped by the instrumental activities of the forums.

In conclusion, this topical landscape reveals the complex interplay of marketplace dynamics, security concerns, social interactions, and illicit commerce that characterize these online communities.

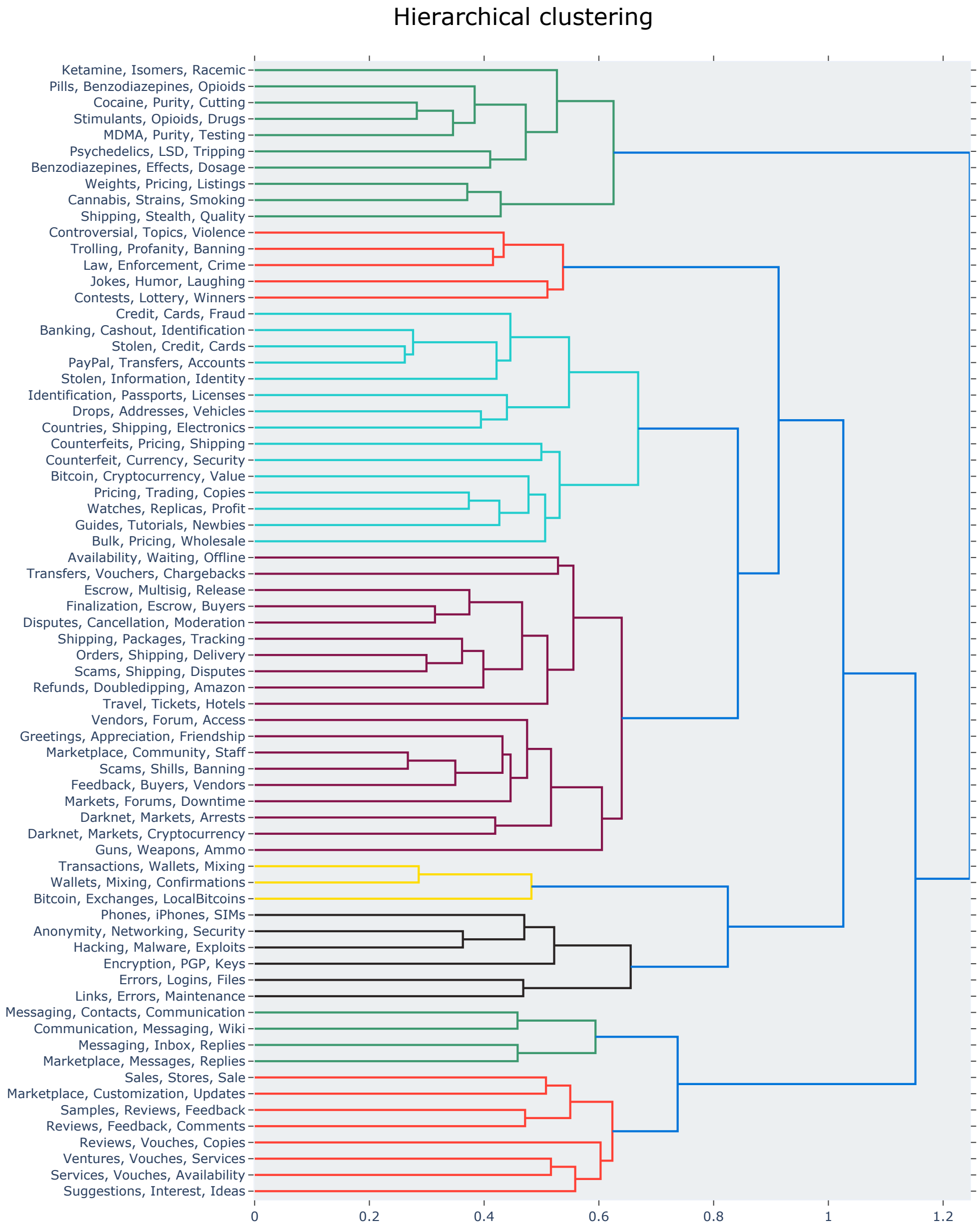


Figure 6.6: Hierarchical clustering of topics. The dendrogram illustrates relationships and distances between topics, with color-coding indicating similarity in content.

These results offer a foundation for further exploring community structure and dynamics through more in-depth analysis of topic co-occurrence patterns, user engagement with different topic clusters, and the network configurations that emerge from these interactions.

6.1.3 Summary of topical landscape analysis

Our topic modelling analysis revealed 70 distinct topics within the Evolution dark web forum, ranging from illicit market transactions and drug trading to security concerns and social interactions. The prevalence of an "Outliers" category (34.17% of posts) highlights the complexity of discussions in these forums. Key topic clusters emerged around marketplace dynamics, security and anonymity, and social aspects of the community. The relationships between these topics, as visualized in our similarity matrix and hierarchical clustering, provide valuable insights into the structure and dynamics of discussions within this dark web forum.

6.2 Combining dimensions of interests and interactions

Having determined the topical landscape, we aim to answer the question: *How does the topical landscape relate to the structural landscape, i.e., the structural communities of the dark web forum communication network?* Understanding the interplay between content and structure is crucial for mapping out the dynamics of discussion forums on the dark web. By integrating topic modelling results with network analysis techniques, we explore how thematic interests align with structural communities.

6.2.1 Structural landscape analysis

To understand the relationship between the topical landscape and the structural landscape of the dark web forum communication network, the first step is to analyze the structural communities. By examining how users cluster based on their interactions, we can identify the main hubs of activity, which will then be correlated with the discovered topics. Based on the previous analysis, we consider giant component with aggregated normalized weights. The results of the analysis are presented in Figure 6.7.

Communities of interaction metrics The analysis of the network subgraphs created using the Leiden algorithm, as visualized in Figure 6.7, reveals significant variances in the structure and connectivity of the communities within the dark web forum communities of interaction. This analysis provides insights into the scale and flow of information among users grouped by their interactions.

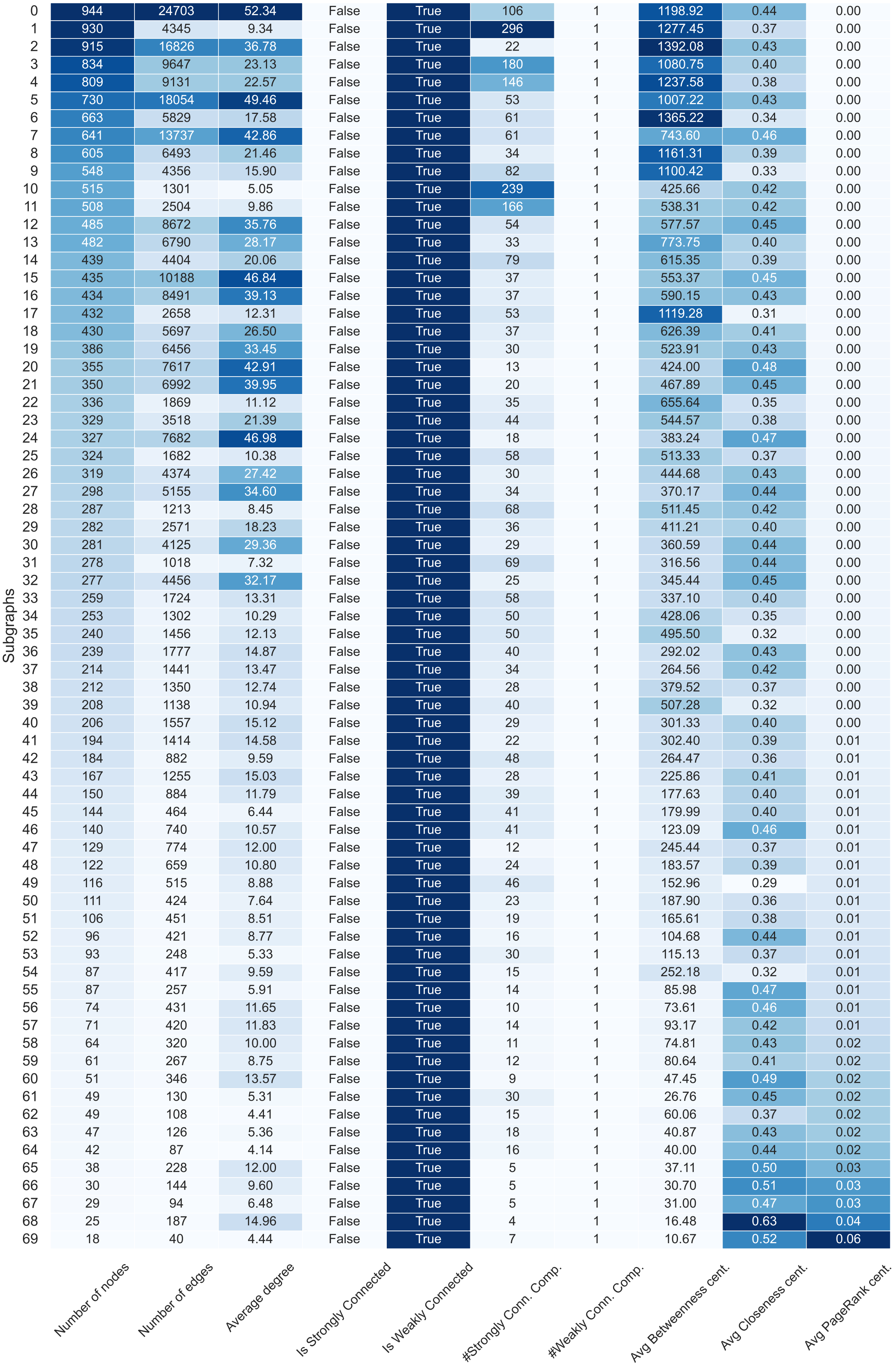
Community size variation The sizes of communities vary significantly, with the largest community containing 24,703 edges and the smallest only 40. This wide disparity suggests that some user groups dominate the discussion more than others, indicating a highly uneven distribution. Edge counts within communities vary in accordance with their size. The largest community, with the most nodes, has 944 nodes, in contrast, the smallest community has only 18 nodes.

Density and connectivity Tracking connectivity within communities reveals significant variation in the average degree, or the average number of connections per node, ranging from 52.34 in the largest community to just 4.44 in the smallest. This suggests that larger communities include more members and promote more interconnectivity among them.

Strong and weak connections An analysis of strong and weak connectivities shows that although all communities are weakly connected, meaning there is some level of connectivity between subgroup components, strong connectivity—where every node can directly reach every other node—is rare. This indicates the presence of directional communication barriers in most communities.

Component analysis The analysis of connected components within each community provides further details. Most communities include multiple strongly connected components, but typically only one weakly connected component. This pattern underscores the presence of central core groups that potentially guide or influence the broader community's discussions.

This detailed analysis of community interactions within the dark web forum highlights the complexity of these relationships. Significant variations in community size, connectivity, and influence underscore the diverse roles and levels of influence among different groups within the network.



Subgraphs	Number of nodes	Number of edges	Average degree	Is Strongly Connected	Is Weakly Connected	#Strongly Conn. Comp.	#Weakly Conn. Comp.	Avg Betweenness cent.	Avg Closeness cent.	Avg PageRank cent.
0	944	24703	52.34	False	True	106	1	1198.92	0.44	0.00
1	930	4345	9.34	False	True	296	1	1277.45	0.37	0.00
2	915	16826	36.78	False	True	22	1	1392.08	0.43	0.00
3	834	9647	23.13	False	True	180	1	1080.75	0.40	0.00
4	809	9131	22.57	False	True	146	1	1237.58	0.38	0.00
5	730	18054	49.46	False	True	53	1	1007.22	0.43	0.00
6	663	5829	17.58	False	True	61	1	1365.22	0.34	0.00
7	641	13737	42.86	False	True	61	1	743.60	0.46	0.00
8	605	6493	21.46	False	True	34	1	1161.31	0.39	0.00
9	548	4356	15.90	False	True	82	1	1100.42	0.33	0.00
10	515	1301	5.05	False	True	239	1	425.66	0.42	0.00
11	508	2504	9.86	False	True	166	1	538.31	0.42	0.00
12	485	8672	35.76	False	True	54	1	577.57	0.45	0.00
13	482	6790	28.17	False	True	33	1	773.75	0.40	0.00
14	439	4404	20.06	False	True	79	1	615.35	0.39	0.00
15	435	10188	46.84	False	True	37	1	553.37	0.45	0.00
16	434	8491	39.13	False	True	37	1	590.15	0.43	0.00
17	432	2658	12.31	False	True	53	1	1119.28	0.31	0.00
18	430	5697	26.50	False	True	37	1	626.39	0.41	0.00
19	386	6456	33.45	False	True	30	1	523.91	0.43	0.00
20	355	7617	42.91	False	True	13	1	424.00	0.48	0.00
21	350	6992	39.95	False	True	20	1	467.89	0.45	0.00
22	336	1869	11.12	False	True	35	1	655.64	0.35	0.00
23	329	3518	21.39	False	True	44	1	544.57	0.38	0.00
24	327	7682	46.98	False	True	18	1	383.24	0.47	0.00
25	324	1682	10.38	False	True	58	1	513.33	0.37	0.00
26	319	4374	27.42	False	True	30	1	444.68	0.43	0.00
27	298	5155	34.60	False	True	34	1	370.17	0.44	0.00
28	287	1213	8.45	False	True	68	1	511.45	0.42	0.00
29	282	2571	18.23	False	True	36	1	411.21	0.40	0.00
30	281	4125	29.36	False	True	29	1	360.59	0.44	0.00
31	278	1018	7.32	False	True	69	1	316.56	0.44	0.00
32	277	4456	32.17	False	True	25	1	345.44	0.45	0.00
33	259	1724	13.31	False	True	58	1	337.10	0.40	0.00
34	253	1302	10.29	False	True	50	1	428.06	0.35	0.00
35	240	1456	12.13	False	True	50	1	495.50	0.32	0.00
36	239	1777	14.87	False	True	40	1	292.02	0.43	0.00
37	214	1441	13.47	False	True	34	1	264.56	0.42	0.00
38	212	1350	12.74	False	True	28	1	379.52	0.37	0.00
39	208	1138	10.94	False	True	40	1	507.28	0.32	0.00
40	206	1557	15.12	False	True	29	1	301.33	0.40	0.00
41	194	1414	14.58	False	True	22	1	302.40	0.39	0.01
42	184	882	9.59	False	True	48	1	264.47	0.36	0.01
43	167	1255	15.03	False	True	28	1	225.86	0.41	0.01
44	150	884	11.79	False	True	39	1	177.63	0.40	0.01
45	144	464	6.44	False	True	41	1	179.99	0.40	0.01
46	140	740	10.57	False	True	41	1	123.09	0.46	0.01
47	129	774	12.00	False	True	12	1	245.44	0.37	0.01
48	122	659	10.80	False	True	24	1	183.57	0.39	0.01
49	116	515	8.88	False	True	46	1	152.96	0.29	0.01
50	111	424	7.64	False	True	23	1	187.90	0.36	0.01
51	106	451	8.51	False	True	19	1	165.61	0.38	0.01
52	96	421	8.77	False	True	16	1	104.68	0.44	0.01
53	93	248	5.33	False	True	30	1	115.13	0.37	0.01
54	87	417	9.59	False	True	15	1	252.18	0.32	0.01
55	87	257	5.91	False	True	14	1	85.98	0.47	0.01
56	74	431	11.65	False	True	10	1	73.61	0.46	0.01
57	71	420	11.83	False	True	14	1	93.17	0.42	0.01
58	64	320	10.00	False	True	11	1	74.81	0.43	0.02
59	61	267	8.75	False	True	12	1	80.64	0.41	0.02
60	51	346	13.57	False	True	9	1	47.45	0.49	0.02
61	49	130	5.31	False	True	30	1	26.76	0.45	0.02
62	49	108	4.41	False	True	15	1	60.06	0.37	0.02
63	47	126	5.36	False	True	18	1	40.87	0.43	0.02
64	42	87	4.14	False	True	16	1	40.00	0.44	0.02
65	38	228	12.00	False	True	5	1	37.11	0.50	0.03
66	30	144	9.60	False	True	5	1	30.70	0.51	0.03
67	29	94	6.48	False	True	5	1	31.00	0.47	0.03
68	25	187	14.96	False	True	4	1	16.48	0.63	0.04
69	18	40	4.44	False	True	7	1	10.67	0.52	0.06

Figure 6.7: Communities of interaction metrics

6.2.2 Relationship between topical and structural landscape

To understand the interplay between thematic content and user interactions in these forums, we have utilized visualizations of embeddings with structural clusters (Figure 6.8) and an analysis of topic distributions across communities (Figure 6.9).

We assigned each user a *predominant topic*, defined as the topic in which the user has the most texts, to integrate the dimensions of nodes and topics. While Figure 6.1 presented topic-based colored embeddings, Figure 6.8 shows the same embedding space colored by communities of interaction. This visualization reveals the distribution of interaction communities across the topical space.

A closer examination of Figure 6.8 reveals that certain structural communities prefer specific regions of the embedding space. For instance, Communities 1, 6, and 7 (warm colors) dominate areas associated with financial and fraud-related discussions. In contrast, Communities 2, 3, and 5 (cold blue) have a notable presence in regions related to drug operations and sales. This suggests a correlation between network structure and thematic content. However, it's important to note that no single topic or region is entirely dominated by one structural community.

The analysis of topic distribution in structural communities (Figure 6.9) further supports and elaborates on these observations. Some topics, such as "Stolen, Credit, Cards" and "Orders, Shipping, Delivery", are prevalent across multiple communities, indicating their broad relevance. While most communities contain various topics, some are highly specialized. For example, "Cannabis, Strains, Smoking", "Psychedelics, LSD, Tripping", and "Guns, Weapons, Ammo" are concentrated in specific communities.

Security-related topics like "Anonymity, Networking, Security" and "Encryption, PGP, Keys" appear across numerous communities, suggesting a universal concern for security measures. This aligns with the observation from Figure 6.8 that Community 2 has a strong presence in security-related areas of the embedding space. Moderation-related topics such as "Trolling, Profanity, Banning" and "Scams, Shills, Banning" are present in multiple communities, indicating widespread moderation efforts.

Drug-related discussions are distributed across various communities, highlighting the pervasive nature of this topic. This is reflected in the embedding space visualization, where we see Community 1 dominating drug-related regions but also other communities present in these areas. Topics related to marketplace dynamics (e.g., "Marketplace, Community, Staff", "Feedback, Buyers, Vendors") appear in multiple communities, reflecting the forum's dual role as a discussion platform and marketplace.

This analysis reveals a complex relationship between topical interests and interaction patterns. While some communities form around specific interests, as evidenced by their clustering in certain regions of the embedding space, others are more diverse in their topical composition. This suggests that user interactions are influenced by both shared interests and other factors such as transaction needs or security concerns.

The distribution of structural communities in the embedding space also provides insights into potential bridges between different topic areas. We can observe areas where multiple communities overlap, such as between drug-related topics and marketplace operations. These intersection points could represent key areas of cross-community interaction and information exchange, contributing to the resilience of the overall forum structure.

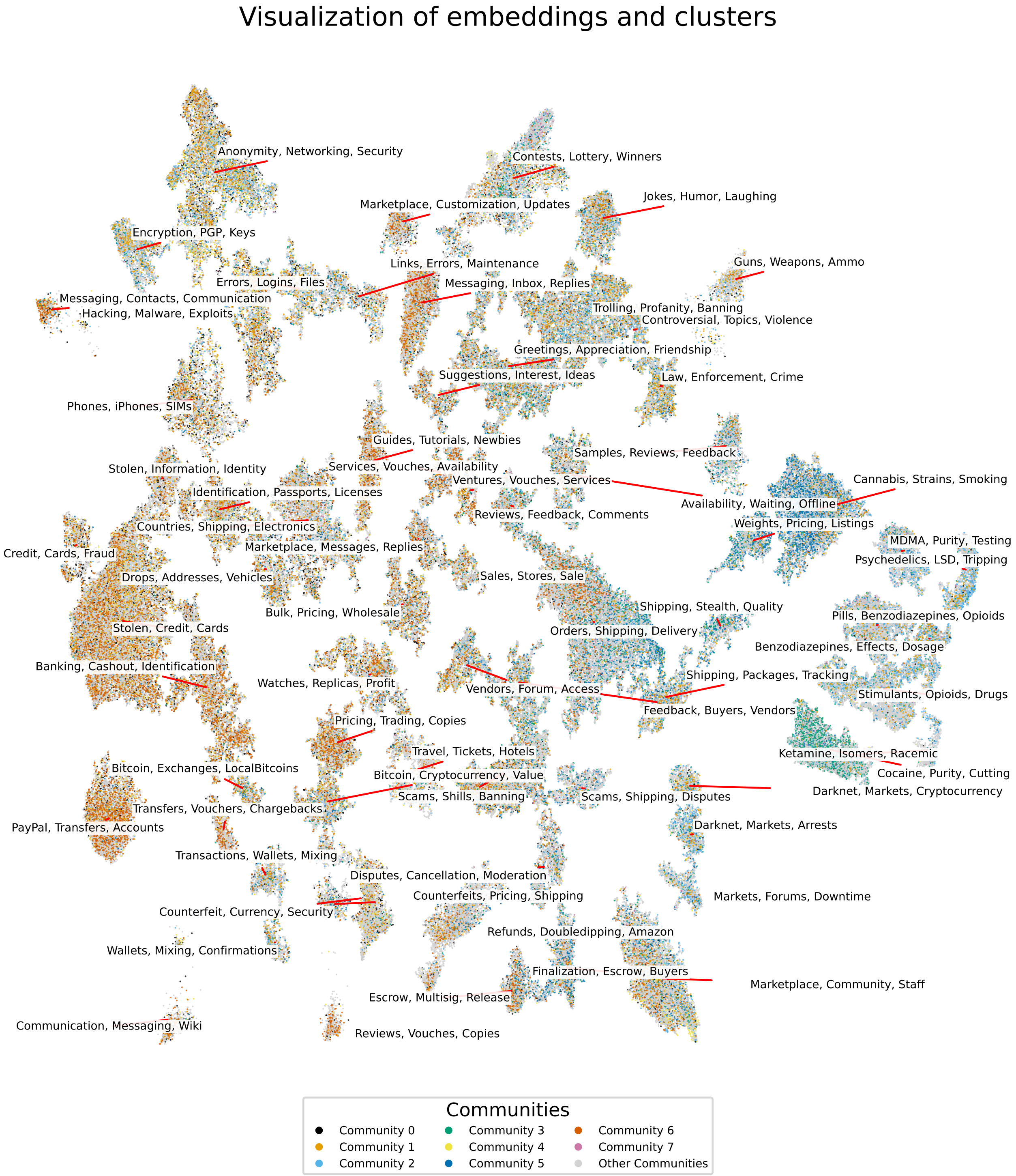


Figure 6.8: Visualization of embeddings and structural clusters. This figure shows how different topics and communities are spread out in the embedded space, with each community color-coded.

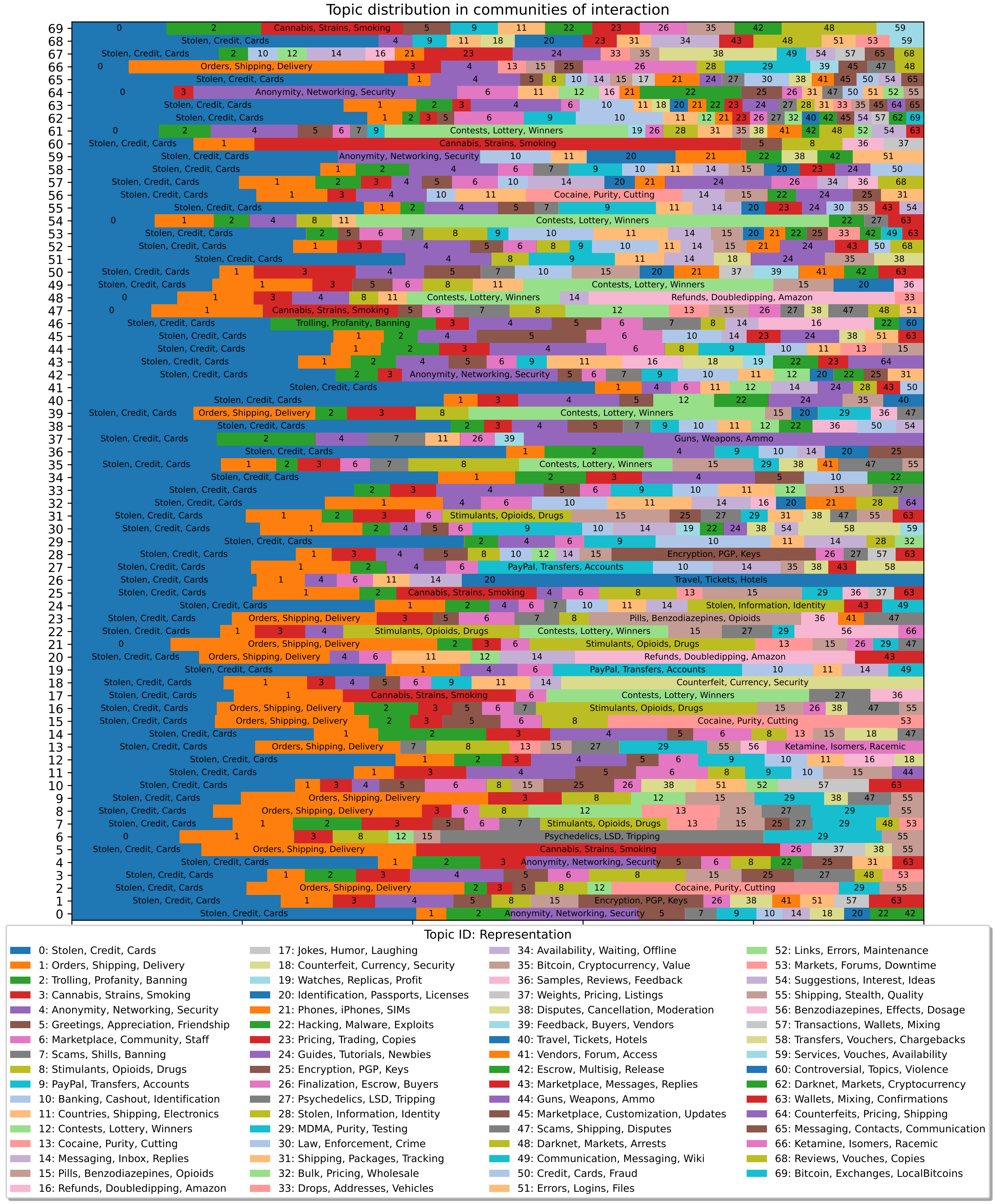


Figure 6.9: Topic distribution in communities of interaction. Each row represents a community with bars indicating the proportion of various topics discussed within that community.

While most communities exhibit a diverse range of topics, our analysis reveals that some communities are highly specialized, focusing on specific areas of interest. For instance, Community 6 is strongly dominated by the topic "Psychedelics, LSD, Tripping", Communities 20 and 48 by "Refunds, Doubledipping, Amazon", Community 26 by "Travel, Tickets, Hotels", Community 38 by "Guns, Weapons, Ammo", Community 54 by "Contests, Lottery, Winners", and Community 60 by "Cannabis, Strains, Smoking". The existence of these specialized communities suggests that certain users gravitate towards specific niches within the dark web ecosystem, forming subgroups with shared interests or expertise. This specialization could indicate the presence of expert users or vendors in particular areas, potentially influencing the dynamics of information flow and transactions within these communities.

Figure 6.10 presents the topic distribution of the top 20 users ranked by betweenness centrality, revealing insights about the most influential users in the network. Notably, the topic "Trolling, Profanity, Banning" is prevalent among 11 of the top 20 users, suggesting that these high-betweenness users often engage in or moderate contentious discussions. This could indicate that these users play a crucial role in managing community dynamics and information flow across different subgroups. Another widely discussed topic among these users is "Anonymity, Networking, Security", highlighting the importance of these issues for key network connectors.

Interestingly, some high-betweenness users focus on niche topics. For example, user 7674's posts are exclusively about "Psychedelics, LSD, Tripping", while user 19161 writes almost entirely on this topic. This specialization suggests that some influential users serve as subject-matter experts, bridging communities with their specific knowledge. Several users (IDs: 81, 2052, 7639) have many posts in "Marketplace, Community, Staff" topics, indicating their potential roles as marketplace facilitators or community leaders. The prevalence of "Stimulants, Opioids, Drugs" in the discussions of users 44, 5674, 14306, 18132, and 18435 further underscores the centrality of drug-related topics in the network.

This analysis of high-betweenness users reveals that they often engage in moderation, security discussions, and specialized topics. Their diverse interests and roles suggest that these users act as important bridges between different communities, facilitating information flow and potentially influencing the structure and dynamics of the dark web forum network.

Our approach not only identifies distinct user groups and their primary activities but also provides valuable insights into the interaction patterns and thematic diversity characterizing these illicit online environments. Future work could build on these findings by developing predictive models to identify emerging threats and key players within such forums.

6.3 Analysis of top-ranked users

In this section, we focus on the relationship between topics and users with high centrality measures scores. Our analysis aims to answer the research question: *"Which topics serve as central hubs among users, and who are the key influencers within these thematic clusters?"* By examining this question, we seek to understand the dynamics of influence and information flow within the dark web forum, identifying both the most prominent discussion themes and the users who play pivotal roles in shaping these discussions.

We approach this analysis through several key visualizations and metrics. Figure 6.11 presents

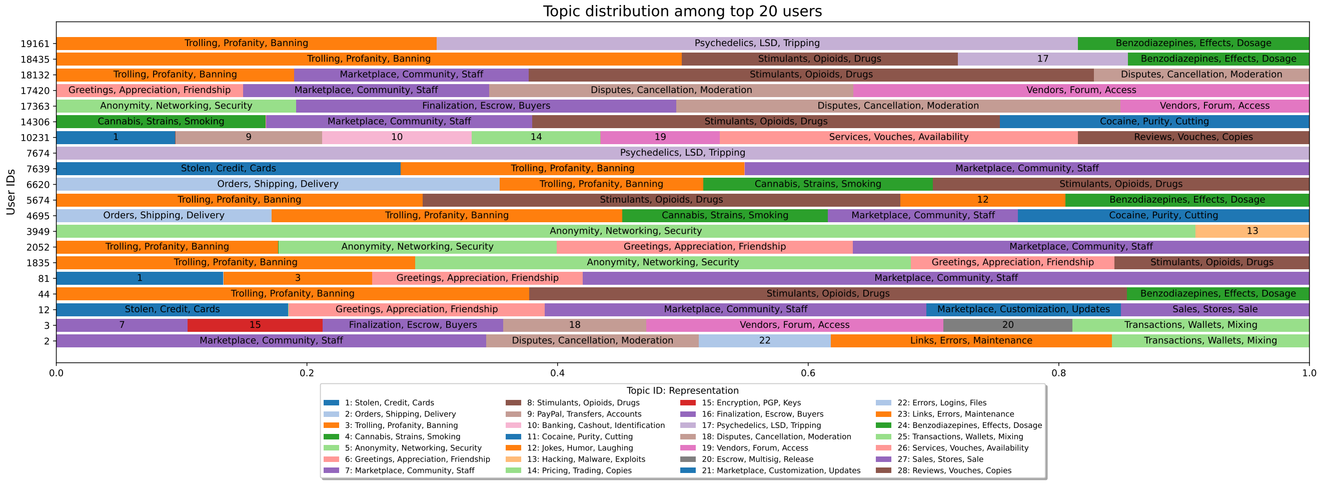


Figure 6.10: Topic distribution of the top 20 influential users in the dark web forum. The figure visualizes the concentration of each user's discussions across various topics, revealing the dominant themes and areas of focus among the most influential members of the community

the average centrality scores by topic, allowing us to identify which themes are most central to the forum's structure. In Figure 6.12, we compare the distribution of topics across top-ranked users with the topic distribution of all nodes, highlighting any differences in thematic focus between influential users and the general forum population. Additionally, Figure 6.13 explores the correlation between centrality measures and sentiment, providing insights into how a user's influence might relate to the emotional tone of their contributions.

Based on Figure 6.11, we can see that having some topic as predominant might determine a high centrality measure. Topic "Trolling, Profanity, Banning", "Counterfeits, Pricing, Shipping" and "Transfers, Vouchers, Chargebacks" show significantly higher values in all 3 out of 4 centrality measures compared to other predominant topics. Only in betweenness centrality scores, we can see some variation. Topic "Vendors, Forum, Access" has a significantly higher score and is not paired with high scores in other centrality measures. "Stimulants, Opioids, Drugs" is the 3rd highest score in betweenness centrality, but its score does not differ substantially with its other scores, keeping its values in the lower half of the distribution.

The analysis of topic distributions raises questions about the exclusivity of predominant topics for top users and the relationship between centrality measures. Figure 6.12 reveals that the topic distribution for the top 20% of users and all nodes is nearly identical, with the most common topics consistent across both groups, differing only in their levels of significance.

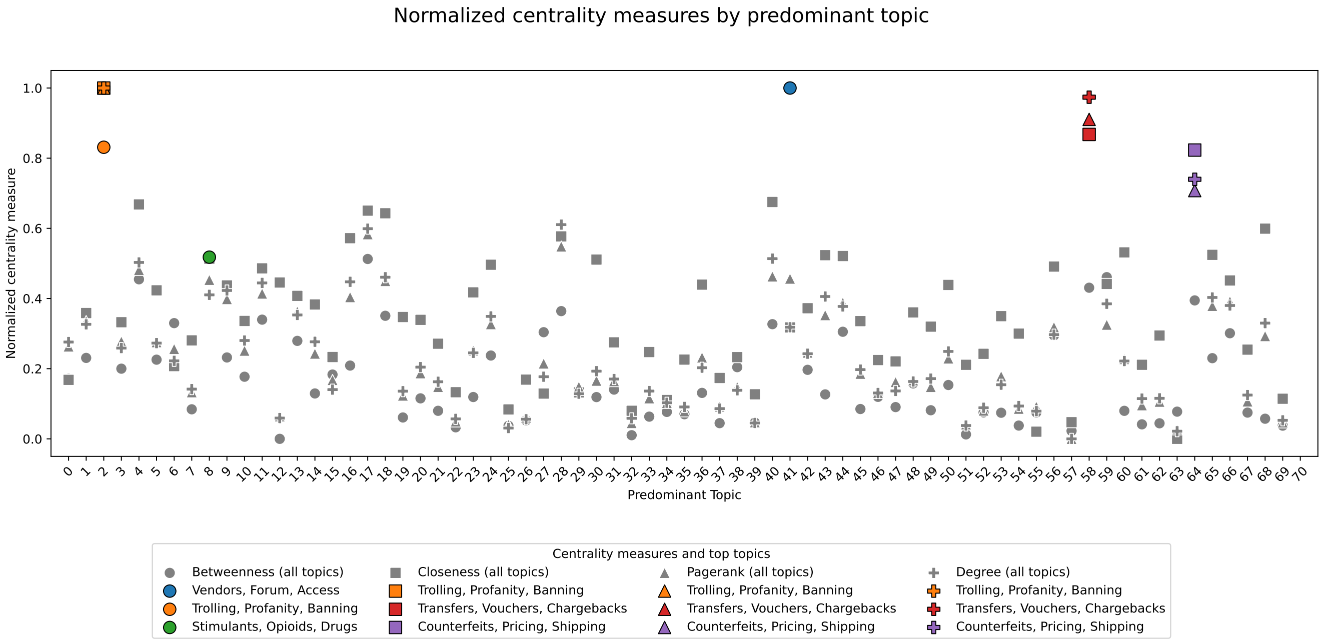


Figure 6.11: Average centrality scores by topic in the dark web forum. The chart presents the topics with the highest average centrality scores, indicating the themes that are most prominently discussed by users holding central positions within the forum's network structure. This analysis highlights the significance of certain topics in shaping the community's interactions and power dynamics.

Topics such as "Orders, Shipping, Delivery", "Anonymity, Networking, Security", "Trolling, Profanity, Banning", and "Cannabis, Strains, Smoking" are frequently discussed by all users but are more prevalent among top users. This prevalence suggests that central users in the network play key roles in facilitating transactions, maintaining security, moderating discussions, and engaging in core marketplace activities. The higher frequency of "Orders, Shipping, Delivery" among top users indicates that they are likely more involved in the logistics and operational aspects of the marketplace. The prominence of "Anonymity, Networking, Security" suggests that these central users are particularly concerned with maintaining the integrity and safety of the network, potentially serving as guardians of the community's security practices. The increased discussion of "Trolling, Profanity, Banning" among top users may indicate that they are more actively involved in community moderation and conflict resolution, shaping the norms and behavior within the forum.

Conversely, topics like "Stolen, Credit, Cards" are commonly discussed by all users but are less frequent among top users. This could suggest that while such illicit activities are widespread in the forum, the most central users may be more focused on facilitating transactions and maintaining the marketplace's infrastructure rather than directly engaging in specific illegal activities.

This pattern of topic distribution indicates that users central to the network tend to be more involved in the operational, security, and community management aspects of the forum, rather than specializing in particular illicit activities. Their role appears to be that of facilitators and moderators, crucial to the functioning and stability of the dark web marketplace ecosystem.

Figure 6.13 presents a heatmap of correlations between centrality measures and sentiment scores for the top-20 nodes and all nodes. The analysis reveals strong correlations among centrality measures for the top-20 nodes, particularly between PageRank and degree. This suggests that the most influential users tend to have both high connectivity and importance as measured by PageRank.

For all nodes, we observe moderate to strong positive correlations between centrality measures and sentiment scores, particularly the subjectivity score. This suggests that nodes with higher centrality might have more subjective content and potentially more positive sentiment. The relationship between sentiment polarity and centrality measures is weaker but still positive.

These correlations offer a nuanced view of network structure and node roles in information flow and community dynamics. The analysis underscores the need for combining centrality measures and sentiment analysis to reveal user influence in dark web forums, as a single metric may miss the network's complexity.

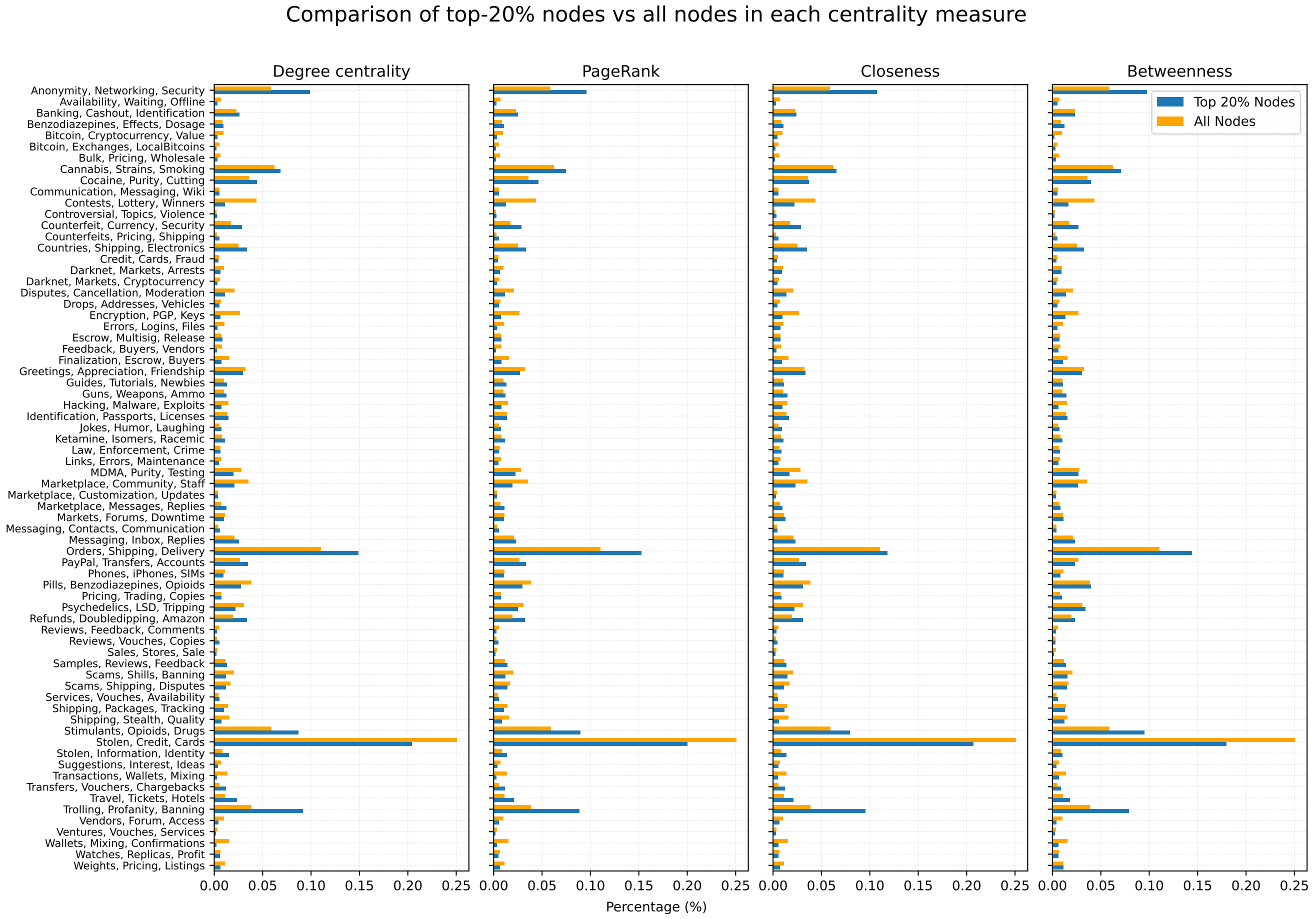


Figure 6.12: Distribution of topics based on centrality measures: A comparison between the top 20% of nodes and all nodes. This figure highlights which topics are more prevalent among the top 20% of users and which are common across all users.

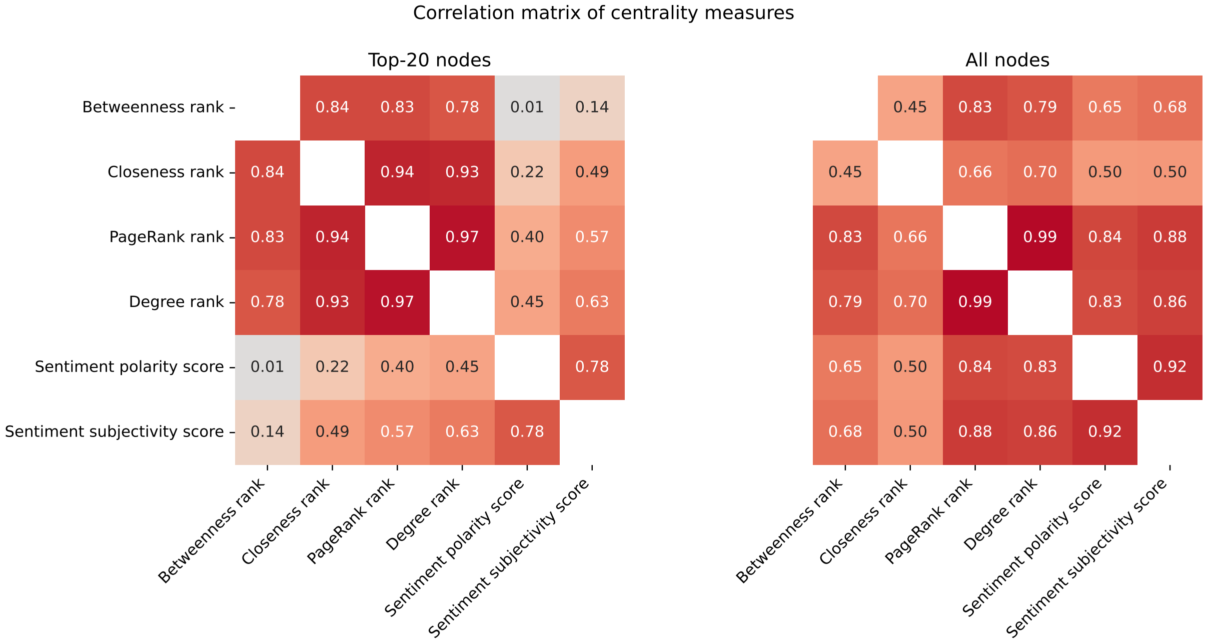


Figure 6.13: Correlation heatmaps for various centrality measures and sentiment scores: This figure demonstrates the correlation between centrality measures and sentiment scores for two distinct node groups: top-20 and all nodes based on betweenness centrality scores.

The sentiment analysis conducted across all users and top users within the dark web forum revealed a striking uniformity in sentiment distribution. Figure 6.14 illustrates that the polarity and subjectivity scores of top users closely mirror the overall user sentiment, with a slight skew towards more objective and less emotional content.

This similarity in sentiment patterns between top users and the general user base suggests that sentiment expressions, whether positive or negative, neutral or subjective, do not vary significantly with a user's centrality or influence within the network. Contrary to what might be expected in a community where persuasive or impactful communication could correlate with network position, the emotional or subjective content of posts does not appear to be a distinguishing factor for influential users.

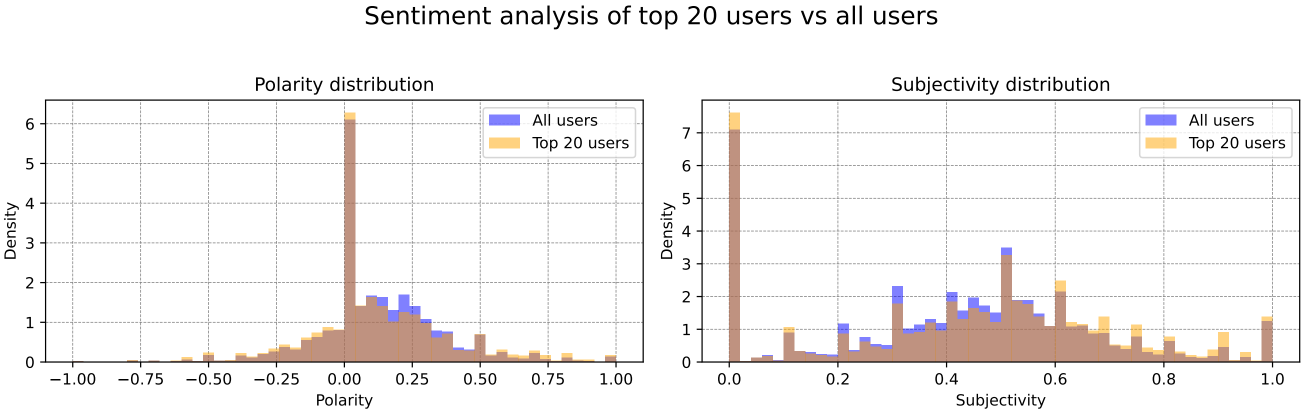


Figure 6.14: Sentiment analysis of top users: The histograms display the distributions of polarity and subjectivity for the top users in comparison to the distribution of all users. These distributions are nearly identical in shape, highlighting a uniform sentiment pattern across all user contributions. The values have been normalized.

The uniformity in sentiment distribution across different user groups, as evidenced by both Figure 6.13 and Figure 6.14, suggests that sentiment analysis may not be a particularly informative tool for understanding user influence or community dynamics in this context. The lack of variation in emotional content between central and peripheral users indicates that factors besides sentiment, such as topic engagement and network position, may be more crucial in determining a user's role and influence within the dark web forum ecosystem.

This finding prompts a reevaluation of the role sentiment analysis plays in understanding network dynamics in dark web forums. It underscores the importance of focusing on network structure and topical analysis, rather than sentiment, when studying the characteristics and dynamics of these online communities. Future research in this area might benefit from exploring alternative methods of content analysis that could potentially reveal more significant differences between influential and general users in these complex networks.

6.4 Global vs. local landscape

In this section, we explore various communities of interaction or structural communities. We aim to analyze the differences between local and global topical landscapes and examine how central users in these communities relate to community topics.

6.4.1 Community 6

To analyze community 6, we prepared Figures 6.15, 6.16, 6.17, and 6.18. These figures present the distribution of centrality scores across predominant topics, the distribution of topics across all nodes and the top 20 nodes in community 6, the topic distribution of the most influential users, and a sentiment analysis, all in scope limited to community 6. These efforts provide valuable insights and allow us to compare this data with the entire dataset.

This community has been chosen for further analysis because of the significantly lower prevalence of the predominant topic across the entire dataset “Stolen, Credit, Cards” and the strongest dominance of topic “Psychedelics, LSD, Tripping” across all structural communities. The community consists of 663 vertices and 5,829 edges, indicating a moderately sized network. With a graph density of 0.0133, it is relatively sparse. The graph’s diameter is 8 and the clustering coefficient 0.3043. The combination of characteristics points to a structure where connections are not overly abundant, but there is a significant degree of local clustering and an interconnected small-world nature.

Based on Figure 6.15, the centrality measures for the topic “Psychedelics, LSD, Tripping” are in the lower half of the spectrum. This is not surprising if most users consider it a predominant topic. Interestingly, topics like “Credit, Cards, Fraud” and “Hacking, Malware, Exploits” stand out in this landscape of psychedelics across all four centrality measures, despite being unrelated to the community’s dominant topic. Other topics correlating with high centrality measures include “Markets, Forums, Downtime” and “Vendors, Forum, Access”. Additionally, the topic “Trolling, Profanity, Banning” also shows high centrality measures.

This pattern of centrality distribution may indicate that while the community itself is highly focused on psychedelics, it still requires highly central users who perform essential functions beyond the primary topic of interest. The prominence of “Credit, Cards, Fraud” and “Hacking, Malware, Exploits” suggests the presence of users with technical expertise who may assist in financial transactions or security measures. The high centrality of “Markets, Forums, Downtime” and “Vendors, Forum, Access” points to users who play crucial roles in maintaining the marketplace’s infrastructure and facilitating transactions.

Moreover, the high centrality of the “Trolling, Profanity, Banning” topic indicates the presence of active moderators within the community. While this topic doesn’t directly relate to psychedelics, it suggests that maintaining community standards and managing conflicts are critical functions performed by central users. This observation underscores the complexity of these online communities, where focused discussion on a specific topic (in this case, psychedelics) coexists with the need for robust community management, technical support, and marketplace facilitation.

Figure 6.16 provides valuable insight into how the predominant topics are distributed in this community, but also shows how exclusive those topics are for the most influential users based on centrality metrics. The figure shows high dominance of topic “Psychedelics, LSD, Tripping”, multiple topics that are prevalent for all users, but top 20 users do not mention like “Cannabis, Strains, Smoking” or “Stolen, Credit, Cards” and topics that are mentioned exclusively by top 20 users like “Trolling, Profanity, Banning”. One interesting observation is the difference in topics mentioned by top users across centrality measures. Degree, PageRank, and Closeness centralities show similar distributions, while Betweenness is notably distinct. For example, the topic “Greetings, Appreciation, Friendship” is mostly mentioned by top 20 users, but only for this triplet of centrality measures. Although there is a difference in the occurrence of certain topics across those measures, when the prevalence is similar, the strength of those topics is mostly the same. For example, ”MDMA, Purity, Testing” has similar strength across all metrics.

The third figure (Figure 6.17) provides a detailed look at the distribution of topics among the top 20 users based on Betweenness centrality in community 6. This figure highlights the dominance of topics such as “Psychedelics, LSD, Tripping” and “Orders, Shipping, Delivery”, along with various other drugs like cannabis, opioids, and MDMA, as well as anonymity-related topics. This distribution indicates that the most influential users in this community are heavily involved in forum operations and discuss various drugs. Although this community is strongly connected to one type of drug trade and experiences, it suggests that other illegal substances are also likely present in the discussions.

The final figures in the analysis of community 6 are the sentiment analysis of this subgroup (Figure 6.18). These figures reveal a semi-uniform distribution of polarity, with a strong dominance of neutral sentiment and a skew towards positive sentiment. The subjectivity scores range from factual to highly emotional, with a dominance of objective sentiment. This distribution closely resembles that of the entire dataset (Figure 4.7), indicating that subcommunities exhibit similar behavior across the forum.

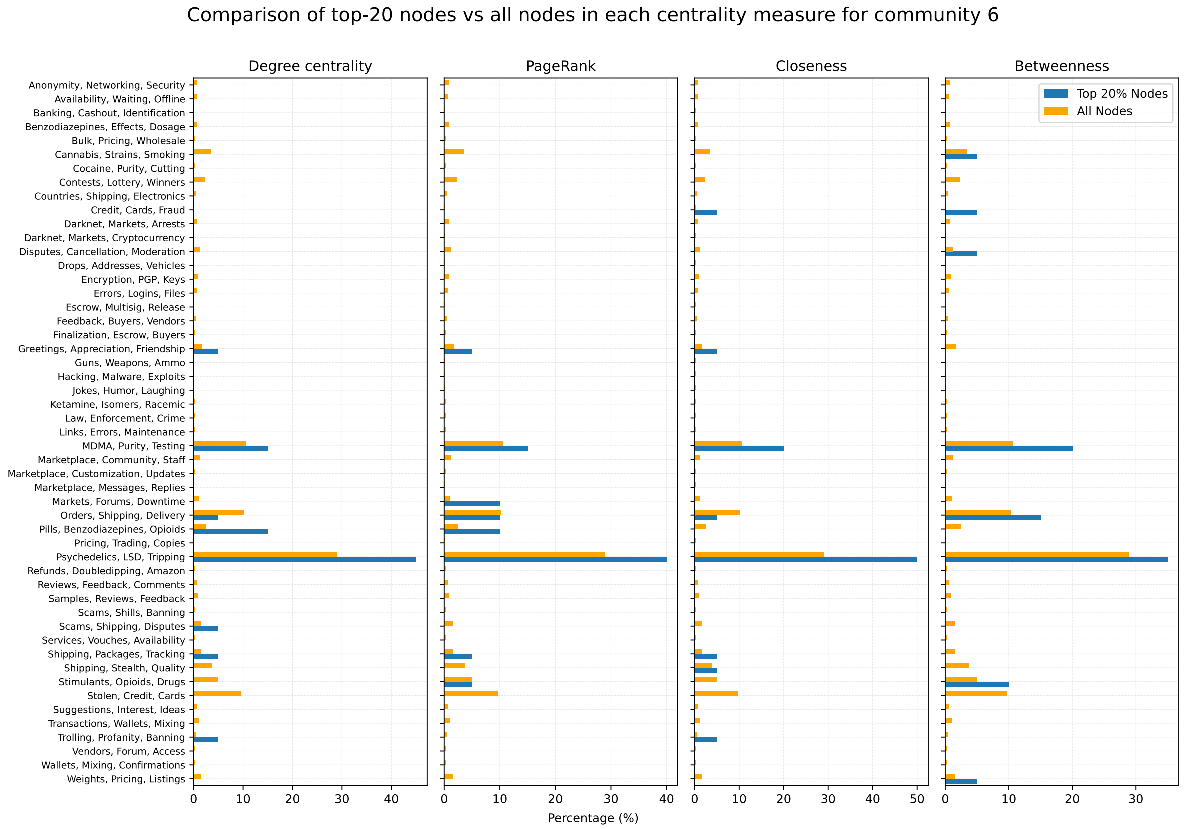


Figure 6.16: Distribution of topics based on centrality measures: A comparison between the top 20 of nodes and all nodes in community 6.

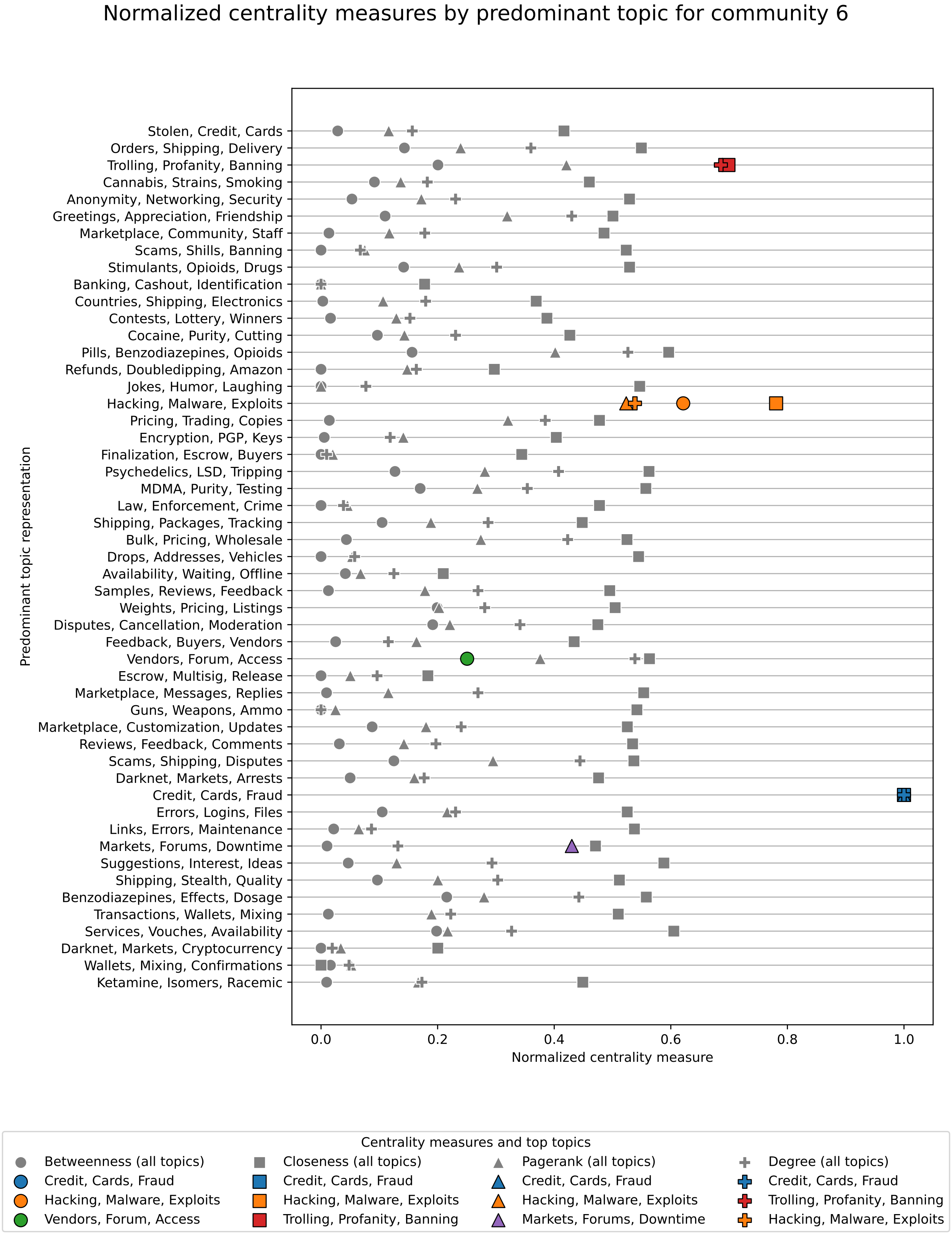


Figure 6.15: Average centrality scores by topic in the community 6.

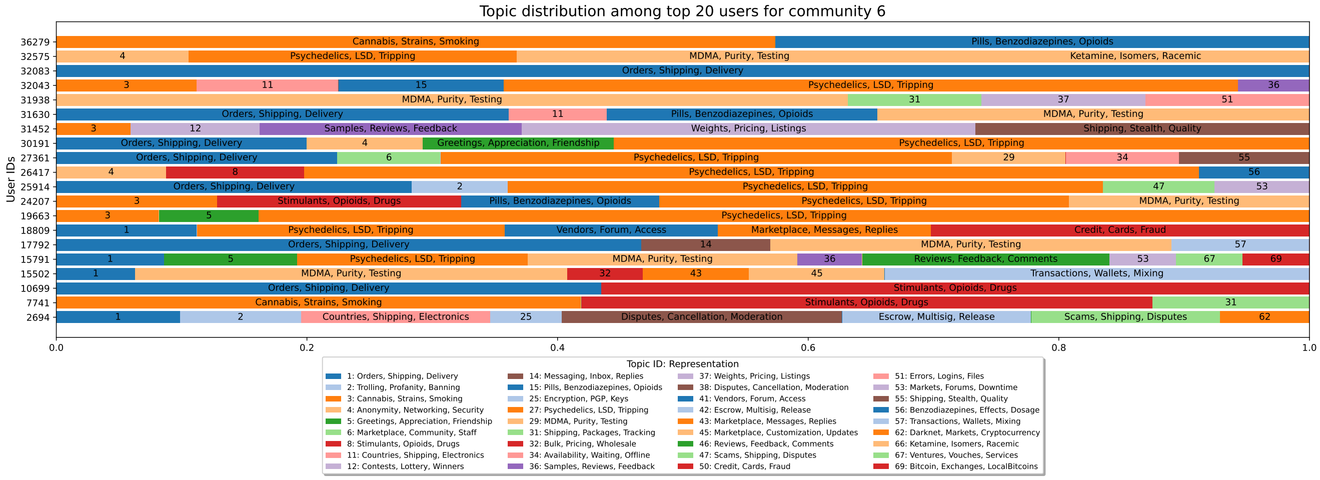


Figure 6.17: Topic distribution of the top 20 influential users in the community 6.

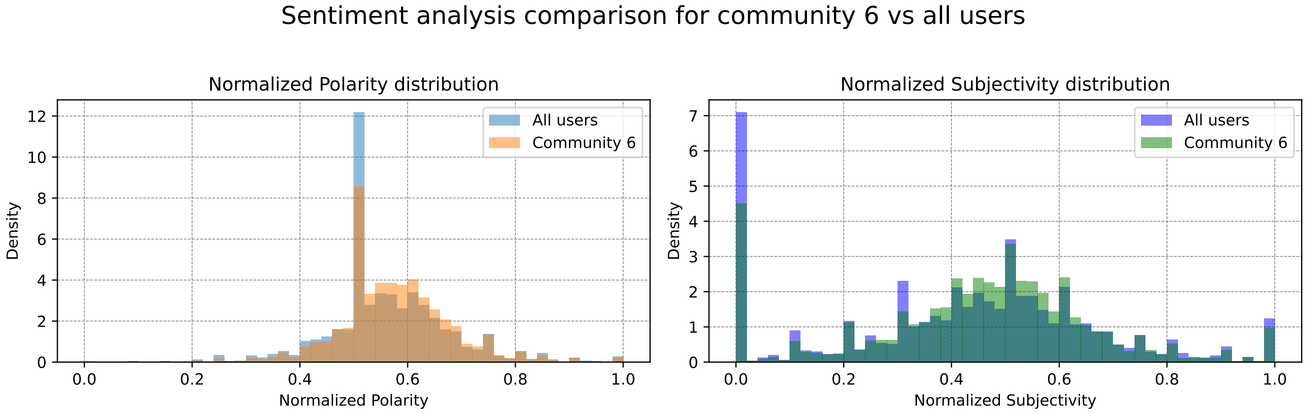


Figure 6.18: Sentiment analysis of community 6 users' posts.

6.4.2 Community 37

Community 37 was selected for further analysis due to its distinct focus on the topic “Guns, Weapons, Ammo”, along with a high prevalence of moderation-related topics such as “Trolling, Profanity, Banning” and “Scams, Shills, Banning”. This community consists of 214 vertices and 1,441 edges, with a density of 0.0316, a diameter of 5, and a clustering coefficient of 0.2354. These metrics suggest a moderately sized network with sparse overall connectivity but significant local clustering and small-world characteristics.

Analysis of centrality measures and topic distributions (Figures 6.19, 6.20, and 6.21) reveals several key insights. The topic “Guns, Weapons, Ammo” clearly dominates across all centrality metrics and user groups, with top users showing up to 80% prevalence in this topic. This strong focus is complemented by the high centrality and prevalence of topics related to forum and marketplace operations, such as “Markets, Forums, Downtime”, “Feedback, Buyers, Vendors”, and “Finalization, Escrow, Buyers”, which consistently appear among top users.

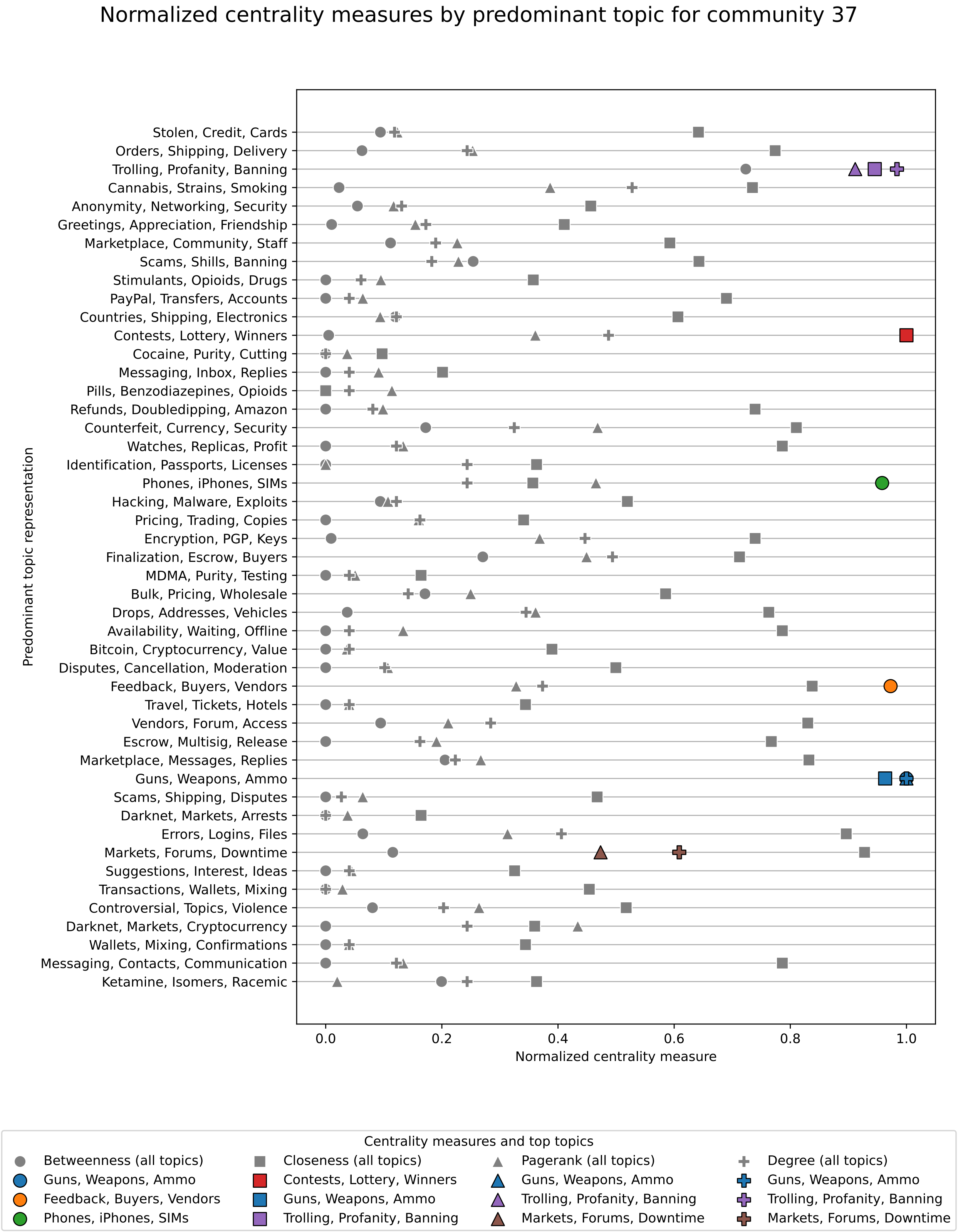
Moderation-related topics, specifically “Trolling, Profanity, Banning” and “Scams, Shills, Banning”, also show high centrality and prevalence among top users, indicating the importance of community management in this group. which aligns with our earlier findings on the global forum level. This focus on operational and moderation topics suggests that while the community is centered around guns and weapons, there is a significant emphasis on maintaining the marketplace’s functionality and community standards.

Interestingly, the analysis also reveals a level of diversity in user interests that goes beyond the dominant theme. Topics such as “Stolen, Credit, Cards” and drug-related subjects like “Stimulants, Opioids, Drugs” are mentioned by all users, following the global trend observed in the overall forum. This suggests that some users in this structural community may have topical interests that don’t align closely with the dominant theme of guns and weapons, highlighting the complex nature of community formation in dark web forums.

A particularly noteworthy observation is the high centrality of the topic “Phones, iPhones, SIMs” (Figure 6.19), which doesn’t obviously relate to gun trade or gun-related topics. This unexpected prominence could indicate diverse interests or activities within the community, or perhaps suggest connections between seemingly unrelated illicit markets.

The sentiment analysis (Figure 6.22) provides an interesting contrast to the topic analysis. Despite the prevalence of potentially contentious topics like “Trolling, Profanity, Banning” and “Scams, Shills, Banning”, the overall sentiment distribution in Community 37 closely mirrors that of the entire forum. This suggests that the dominant topic “Guns, Weapons, Ammo” may have a normalizing effect on the emotional content of posts, resulting in a more neutral overall sentiment, or that discussions about weapons are conducted in a surprisingly neutral tone.

These findings underscore the complex interplay between structural connections and topical interests in dark web forums. While the community is strongly focused on a specific topic, it also reflects broader forum-wide trends and includes users with diverse interests. This complexity challenges simple categorizations and highlights the need for nuanced analysis in understanding dark web community structures. The coexistence of a dominant topic with varied interests, the importance of operational and moderation topics, and the unexpected presence of seemingly unrelated themes all contribute to a multifaceted picture of this community’s dynamics.



t

Figure 6.19: Average centrality scores by topic in the community 37.

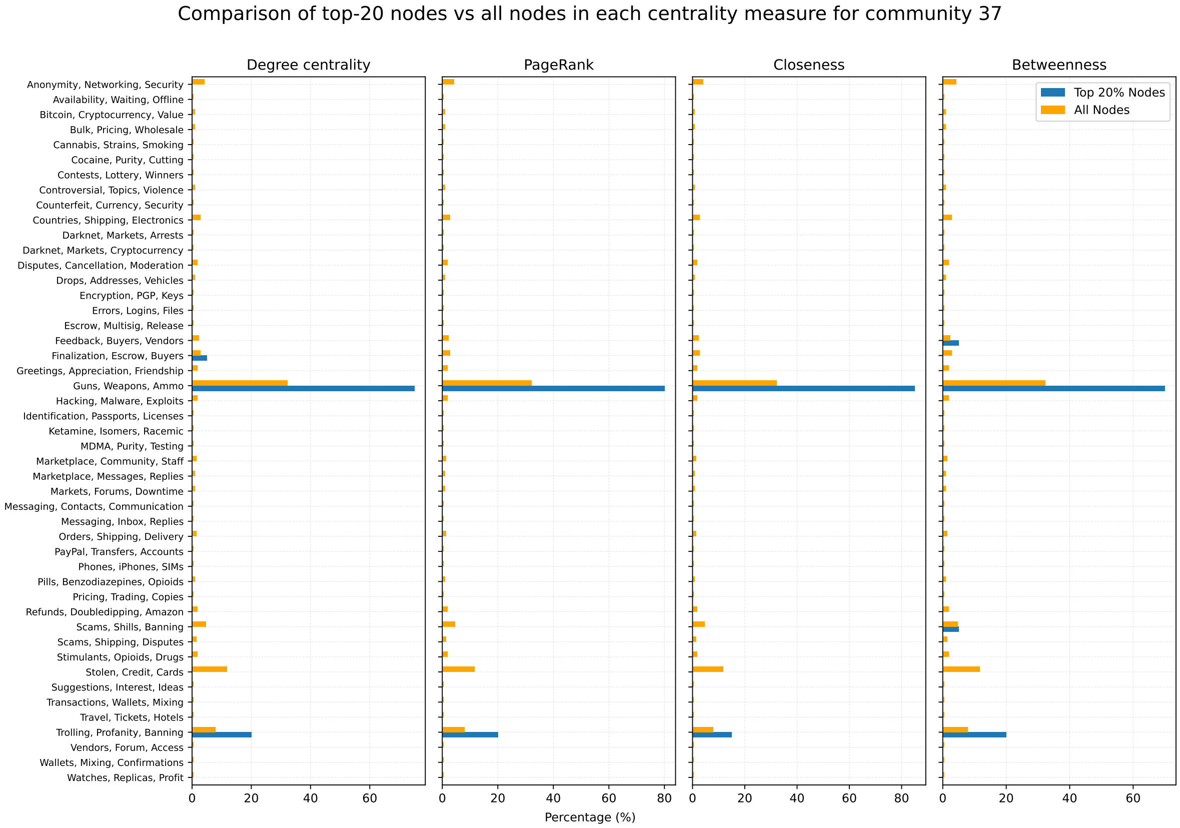


Figure 6.20: Distribution of topics based on centrality measures: A comparison between the top 20 of nodes and all nodes in community 37.

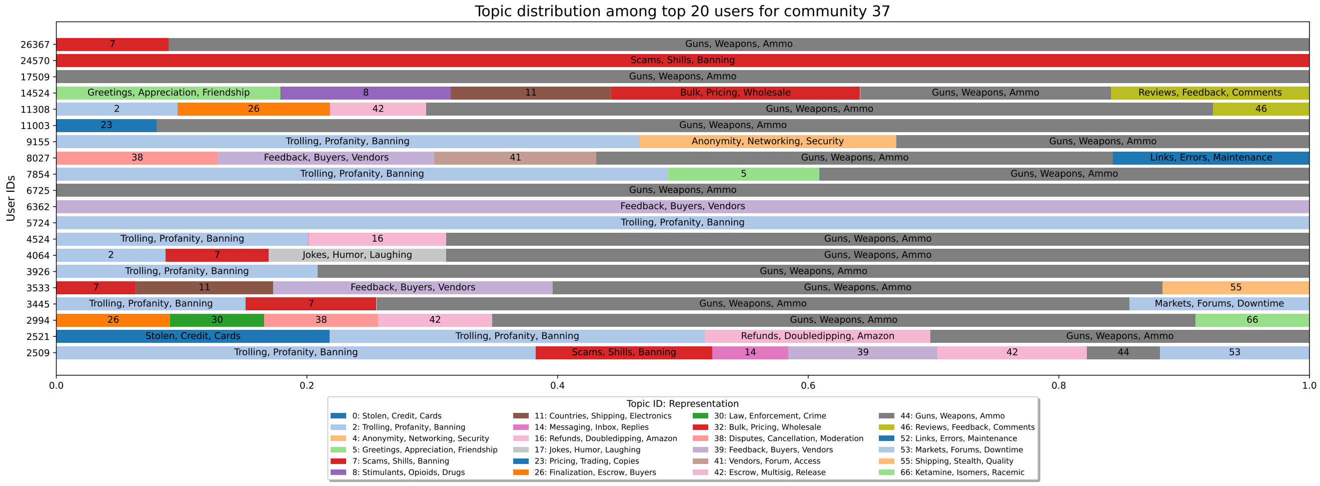


Figure 6.21: Topic distribution of the top 20 influential users in the Community 37.

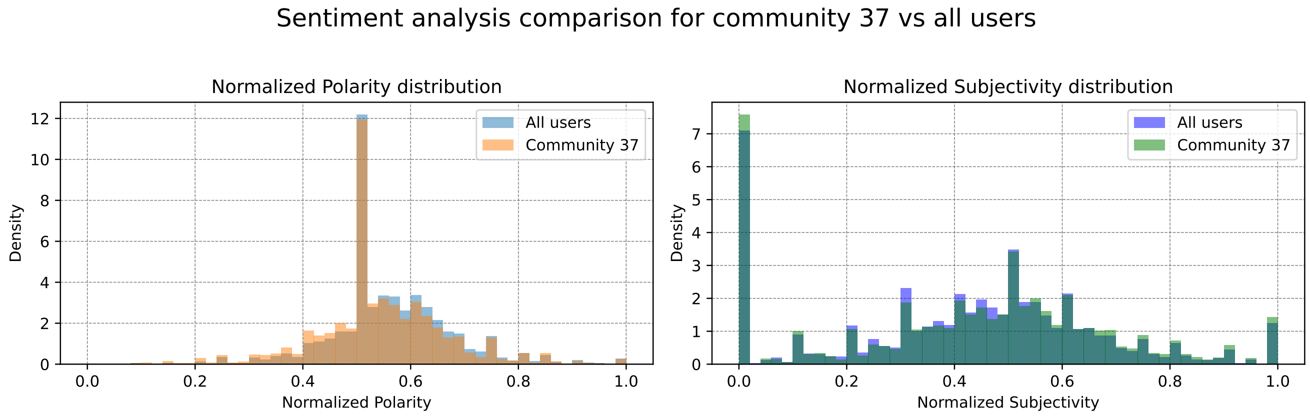


Figure 6.22: Sentiment analysis of Community 37 users' posts.

Chapter 7

Conclusion

This thesis examined community structures in dark web forums using advanced natural language processing and network analysis techniques. Our research uncovered the hidden layers of social interactions and hierarchies in these covert communities associated with various online illicit activities.

Our analysis of the Evolution forum's topical landscape revealed a complex and multifaceted structure. We identified 70 distinct topics, ranging from illicit market transactions and drug trading to security concerns and social interactions. Key topic clusters emerged around marketplace dynamics, security and anonymity, and social aspects of the community. Drug-related topics formed a significant cluster, including discussions on cannabis, stimulants, opioids, and psychedelics. Transactional aspects of dark web marketplaces were prominent, with topics like "Stolen, Credit, Cards" and "Orders, Shipping, Delivery" among the most frequent. Security and trust-related topics, such as "Anonymity, Networking, Security" and "Encryption, PGP, Keys", also featured prominently. Notably, a significant "Outliers" category, accounting for 34.17

In exploring the relationship between topical and structural landscapes, we found that certain structural communities preferred specific regions of the embedding space, suggesting a correlation between network structure and thematic content. While some communities were highly specialized, focusing on topics like psychedelics or guns, others displayed more diverse topical compositions. We identified intersection points between different topic areas and communities, potentially representing key areas of cross-community interaction and information exchange.

Our investigation into central topics and key influencers revealed that topics such as "Trolling, Profanity, Banning", "Counterfeits, Pricing, Shipping", and "Transfers, Vouchers, Chargebacks" showed significantly higher centrality values. The most influential users tended to be more involved in operational, security, and community management aspects rather than specializing in particular illicit activities. High-betweenness users often engage in moderation, security discussions, and specialized topics, acting as important bridges between different communities. Interestingly, sentiment analysis indicated that the emotional content of posts did not appear to be a distinguishing factor for influential users.

Comparing the characteristics of the global dark web network with its individual communities, we found that while individual communities often formed around specific topics, they still reflected broader forum-wide trends. Both global and local networks exhibited sparse connectivity but significant local clustering and small-world characteristics. Even in highly specialized communities, topics related to marketplace operations and security remained important among influential users. Local communities exhibited similar sentiment patterns to the global network, with a prevalence of neutral sentiment and a slight positive skew.

In examining how the characteristics of the global dark web network compare to those of its individual communities, our analysis revealed a complex interplay between specialized interests and broader forum dynamics. While individual communities often formed around specific topics such as psychedelics or weapons, they still reflected broader forum-wide trends. Both global and local networks exhibited sparse connectivity but significant local clustering and small-world characteristics. Notably, even in highly specialized communities, we observed the presence of central users who perform essential functions beyond the primary topic of interest, facilitating broader forum functions, maintaining security, or providing cross-topic expertise. This pattern underscores the need for nuanced analysis in understanding dark web community structures, highlighting how local landscapes can differ from yet still mirror global forum trends.

These findings contribute to our understanding of dark web social structures, offering insights into influence and community formation in these hidden spaces. They underscore the importance of combining content analysis with network structure examination when studying complex online ecosystems. While the methodologies developed here may be useful for analyzing hidden online communities, their applicability beyond the dark web context requires further investigation. As with all research in this field, our findings should be interpreted cautiously, recognizing the complex and ever-changing nature of dark web forums.

Bibliography

Alash, H. M. and Al-Sultany, G. A. (2021). Enhanced twitter community detection using node content and attributes. In *2021 1st Babylon International Conference on Information Technology and Science (BICITS)*, pages 5–10.

Barabási, A.-L. (2013). Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375.

Basheer, R. and Alkhatib, B. (2024). Conceptualizing discussions on the dark web: An empirical topic modeling approach. *Complexity*, 2024:1–24.

Bavelas, A. (1950). Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, 22(6):725–730.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Boekhout, H. D., Blokland, A. A., and Takes, F. W. (2023). A large-scale longitudinal structured dataset of the dark web cryptomarket evolution (2014-2015). arXiv:2311.11878 [cs.SI].

Boekhout, H. D., Blokland, A. A. J., and Takes, F. W. (2024). Early warning signals for predicting cryptomarket vendor success using dark net forum networks. *Scientific Reports*, 14(1):16336.

Branwen, G., Christin, N., Décary-Hétu, D., Andersen, R. M., StExo, Presidente, E., Anonymous, Lau, D., Sohhlz, Kratunov, D., Cakic, V., Buskirk, V., Whom, McKenna, M., and Goode, S. (2015). Dark net market archives, 2011-2015. `https://gwern.net/dnm-archive`. Dataset. Accessed: 2024-08-18.

Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

Chen, H. (2008). Sentiment and affect analysis of dark web forums: Measuring radicalization on the internet. In *2008 IEEE International Conference on Intelligence and Security Informatics*, pages 104–109.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

DarkOwl (2024). Darkowl glossary of darknet terms. `https://www.darkowl.com/resources/darkowl-glossary-of-darknet-terms/`. Accessed: 2024-05-31.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining*.

Fonhof, A. M. P., van der Bruggen, M., and Takes, F. W. (2019). Characterizing key players in child exploitation networks on the dark net. In Aiello, L. M., Cherifi, C., Cherifi, H., Lambiotte, R., Lió, P., and Rocha, L. M., editors, *Complex Networks and Their Applications VII*, pages 412–423, Cham. Springer International Publishing.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv:2203.05794 [cs.CL].

Grootendorst, M. (2023). Bertopic. `https://maartengr.github.io/BERTopic/index.html`. Online documentation. Accessed: 2024-08-18.

Halkidi, M. and Vazirgiannis, M. (2008). A density-based cluster validity approach using multi-representatives. *Pattern Recognition Letters*, 29:773–786.

Han, Z., Li, S., Cui, C., Han, D., and Song, H. (2019). Geosocial media as a proxy for security: A review. *IEEE Access*, 7:154224–154238.

IACA (2024). Iaca dark web dictionary. `https://iaca-darkweb-tools.com/dictionary/`. Accessed: 2024-05-31.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2018). Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. arXiv:1711.04305 [cs.IR].

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv:1607.01759 [cs.SI].

Kanavos, A., Perikos, I., Hatzilygeroudis, I., and Tsakalidis, A. (2017). Emotional community detection in social networks. *Computers Electrical Engineering*, 65.

Kostelac, M. (2021). Comparison of language identification models. `https://modelpredict.com/language-identification-survey`. Accessed: 2024-08-18.

Loria, S. (2024). Textblob: Simplified text processing. `https://textblob.readthedocs.io/en/dev/`. TextBlob Documentation. Accessed: 2024-08-18.

McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2.

McInnes, L., Healy, J., and Melville, J. (2020). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 [stat.ML].

Nolker, R. and Zhou, L. (2005). Social computing and weighting to identify member roles in online communities. In *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pages 87–93.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pete, I., Hughes, J., Chua, Y. T., and Bada, M. (2020). A social network analysis and comparison of six dark web forums. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroSPW)*, pages 484–493.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2018). Language models are unsupervised multitask learners.

Raschka, S., Patterson, J., and Nolet, C. (2020). Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence.

Richardson, L. (2024). Beautiful soup documentation. `https://www.crummy.com/software/BeautifulSoup/bs4/doc/`. Online documentation. Accessed: 2024-08-18.

Soska, K. and Christin, N. (2015). Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 33–48, Washington, D.C. USENIX Association.

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1).

Van Kooten, P. (2024). textsearch: A Library for Comparing Text in Python. `https://github.com/kootenpv/textsearch`. GitHub repository. Accessed: 2024-08-18.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. arXiv:1910.03771 [cs.SL].

Xu, J. and Chen, H. (2008). The topology of dark networks. *Communications of the ACM*, 51(10):58–65.

Yang, J., Ye, H., and Zou, F. (2020). pydnettopic: A framework for uncovering what darknet market users talking about. In Park, N., Sun, K., Foresti, S., Butler, K., and Saxena, N., editors, *Security and Privacy in Communication Networks*, pages 118–139, Cham. Springer International Publishing.

Yang, L., Liu, F., Kizza, J. M., and Ege, R. K. (2009). Discovering topics from dark websites. In *2009 IEEE Symposium on Computational Intelligence in Cyber Security*, pages 175–179.