A Social Network Analysis and Comparison of Six Dark Web Forums

Ildiko Pete, Jack Hughes, Yi Ting Chua, Maria Bada
Department of Computer Science and Technology, University of Cambridge
Cambridge, UK
firstname.lastname@cl.cam.ac.uk

***Abstract*—With increasing monitoring and regulation by platforms, communities with criminal interests are moving to the dark web, which hosts content ranging from whistle-blowing and privacy, to drugs, terrorism, and hacking. Using post discussion data from six dark web forums we construct six interaction graphs and use social network analysis tools to study these underground communities. We observe the structure of each network to highlight structural patterns and identify nodes of importance through network centrality analysis. Our findings suggest that in the majority of the forums some members are highly connected and form hubs, while most members have a lower number of connections. When examining the posting activities of central nodes we found that most of the central nodes post in sub-forums with broader topics, such as general discussions and tutorials. These members play different roles in the different forums, and within each forum we identified diverse user profiles.**

***Index Terms*—Social Network Analysis, Online communities, Dark Web, Social Interactions**

1. Introduction

Social networks within online underground forums have received research attention with interests ranging from marketplaces [1] [2], digital underground economies [3] to analysing key actors [4]. At the core of these underground forums are members who interact with each other. Social network analysis provides valuable insights to how these communities operate. The structures, operations, and interactions of dark web communities are still largely undiscovered due to difficulties associated with data collection. Therefore, we direct our attention to a specific subset of these communities, which operate exclusively on the dark web. The dark web refers to a section of the Internet which can only be accessed through anonymity tools and networks, such as Tor[1], and is not indexed by search engines. The dark web has come to be known primarily for enabling anonymity while engaging in illegal activities. Hidden services host a wide array of content, including whistle-blowing, privacy, drugs, terrorism, and hacking [5]. Previous research on dark web forums has mostly focused on terrorist content and extremist groups [5] [6]. However, it presents a rich ecosystem of communities and research efforts need to include other dark web communities with other criminal interests. This is a growing concern as cryptomarkets for drugs, firearms, and cybercrime begin to migrate towards the dark web.

We perform an analysis and comparison of six dark web cybercrime-related forums of different languages, to provide insights into their organisational structures that underlie information and resource flows within these communities. Specifically, we:

- Construct six weighted undirected networks to model interactions between members
- Observe the structure of each network, to highlight structural characteristics and patterns
- Identify nodes of importance in each community using network centrality analysis

Through these analyses we seek to answer the following questions:

- What large scale structural attributes do these networks possess and how do these affect interactions taking place on these forums?
- What are the posting activities of central nodes in these networks?

It is important to note that these networks evolve, however this study focuses on a static view of interactions. We use a snapshot of posts between *February 2019* and *July 2019* inclusive, selected as preliminary analysis found this period to have the highest interaction activity within our dataset. Also, it is not within the scope of this exploratory work to provide a detailed analysis supporting disruption strategies. However, in carrying out this study, our findings contribute to a deeper understanding of these communities, which may lay the foundation for future research on intervention or disruption activities and techniques.

The paper is structured as follows. First, we present related work in Section 2, followed by a description of the data used for analysis in Section 3, and the network model in Section 4. We present the methods of the analysis in Section 5, and provide results in Section 6, which are discussed in Section 7.

2. Related Work

Users can interact in a wide variety of online underground forums, which either operate on the surface web, or on the dark web. The social networks within forums on the surface web have been widely researched from various angles, such as analysing the hyperlinks networks between websites and blogs on child exploitation [7], the social networks across carding forums [8], social networks of malware writers and hackers [9], [10], or systematically

1. https://www.torproject.org/

identifying key hackers for keylogging tools within a large English–language hacker forum [11].

Previous work related to the network analysis of the dark web focus on two main areas. The first area is concerned with investigating the structure of the dark web. For example, Griffith et al. [12] built a network of onion domains and performed network analysis on them. Similarly, in another study [13] the authors constructed a network using Tor hyperlinks and applied social network measures to reveal the network structures that form on Tor. Another work [14] focused on the structural analysis of the dark web through its topological representation. The authors examined how the structural properties including network size, average path length, and the global clustering coefficient of the dark web change over time.

The second larger area of previous work within dark web social network analysis studies dark web forums, which is the focus of our study. Phillips et al. [15] analyse the social structure of dark web forums collected as part of the Dark Web Forum Portal dataset aiming to identify potentially 'important' members of Islamic Networks within these forums. Da Cunha et al. [16] examine a child pornography ring acting inside the dark web and identified that the core of strongly connected criminals lacks a modular structure unlike typical criminal networks. The work closest to ours is by Zamani et al. [17] who utilise the structural properties of two types of interaction networks constructed from various dark web and underground forums. The authors compared these forums and identified that they can be categorised as public, semi-dark and dark web forums. The authors also take into account the evolving nature of networks and examine their dynamics. Our study differs from this work as we do not aim to reveal differences between dark, semi-dark and public forums through analysing their structural properties. We analyse structural properties of multiple dark web forums to gain insights to how connected and centralised these networks are. A novel contribution of our study is that we carry out an analysis of nodes we identify to be influential based on their centrality scores, and aim to understand the roles they play on the forums through a qualitative analysis of their posting activities.

3. The CrimeBB dataset

The data used in this work is from the CrimeBB dataset, provided by the Cambridge Cybercrime Centre (CCC)[2] [18]. CrimeBB is a collection of public forum data scraped from various dark and surface web forums, and it is available through a data sharing agreement with the CCC for researchers who wish to analyse underground cybercriminal communities. Each forum in CrimeBB contains sub-forums, which are organised around specific subjects. Some sub-forums serve as marketplaces, while others are platforms for discussing specific topics and share knowledge in the form of posts. Sub-forums each contain a number of threads: an ordered set of posts, with the first post setting the topic of a thread, and later posts replying to or discussing this.

Selected Forums. CrimeBB contains data from 25 forums and discussion platforms. Of these, we selected six, as they are hosted on hidden services, which is the focus of our analysis. These include English, German and Russian language forums, which we anonymise to prevent unintended consequences, by assigning a label to each forum to be able to refer to them throughout the paper. The scale of data used in this analysis is shown in Table 1. The largest forum based on the number of posts is Forum C, followed by Forum D. Both possess a considerably larger number of authors compared to the others.

Forum	Number of Posts	Number of Authors	First post date
Forum A	88,753	8,242	2014.01.09
Forum B	12,424	1,127	2018.11.22
Forum C	249,880	40,763	2018.02.15
Forum D	240,628	17,241	2012.01.11
Forum E	59,365	8,231	2013.10.23
Forum F	28,485	3,812	2017.05.25

TABLE 1. SELECTED FORUMS

Forum A is one of the most popular dark web general purpose discussion forums. Our dataset contains 62 sub-forums including markets, discussions on security, Tor, drugs, and cryptocurrencies.

Forum B is a forum with a focus on free speech. Our dataset features boards discussing cryptocurrencies, freedom of speech, website hosting, and software.

Forum C is a major dark web forum. Its sub-forums feature discussions on privacy, the dark web, and hacking.

Forum D is another major dark web forum, which provides a platform to discuss a wide range of topics including cryptocurrencies, politics, weapons, psychology, hacking, and privacy.

Forum E offers discussions on torrents in addition to general hacking.

Forum F members can discuss topics related to malware, social engineering, cryptography, and hardware, amongst others.

Ethical Considerations. We received approval for this work from the ethics committee of the Department of Computer Science and Technology. The ethics approval was granted after the review of our research project based on criteria such as data handling, confidentiality and anonymity, risks to participants and to researchers. The data collected from these dark web forums are publicly available. We do not publish any usernames or process personal information. In accordance with the British Society of Criminology's Statement on Ethics, this approach is justified as the dataset is collected from public forums. Therefore, no consent was obtained from users, as this would be infeasible and contradictory to the goals of our study. Our research takes place on collective behaviour, without aiming to identify particular members.

4. Network Model

All of these forums share a similar *hierarchy*, each containing a fixed series of *subforums* based on general topics chosen by the forum administrators, which in turn are composed of member-contributed *threads* based around a single topic. Threads contain an ordered collection of *posts*, which can be either replies to the *author* of the first post, replies to other post authors, or general comments. Longer, more general threads containing many posts may go "off-topic", as the discussion changes focus

2. https://www.cambridgecybercrime.uk

away from the original topic. Thus, forum interactions can take various forms ranging from simple scenarios to increasingly complex ones spanning a larger time period with gaps between active post periods. Our network model, shown by Figure 1, is based upon this hierarchy and models public *interactions* between participants of underground forum discussions.

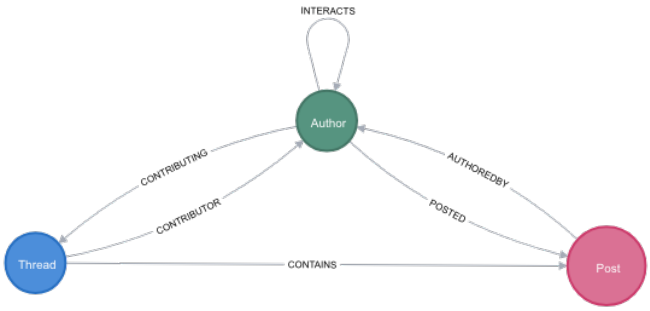

Figure 1. Conceptual Network Model with All Entity Types.

4.1. Nodes

Based on forum structure we derive three conceptual entities that map to nodes in our model. Participants of forum discussions, which we refer to as 'Authors', represent one node type. 'Authors' discuss subjects in the form of 'Posts', which we also model as a network node. Since 'Posts' can be grouped into 'Threads', the latter forms the third conceptual entity of the model.

4.2. Edges

Based on forum interactions we define an *interaction relationship* between two authors if both authors post in the same thread. Authors who post in multiple threads together will have multiple interaction relationships connecting them, which allows weights to be associated with these edge types. As shown by Figure 1, there are three further edge types (*Author-Post, Thread-Post, Author-Thread*) present in our schema. These are not directly used in our analyses but support it. For example, they are used to link authors to threads to identify topics a given member has interest in, or authors to posts to allow reporting on posting behaviour.

4.3. Interaction Network

Interaction relationships between 'Author' nodes (forum members) form the 'Interaction Network', which all network analysis presented in Section 5 is carried out on.

Interactions are initiated by the original thread poster. Any member replying to the thread at any given time will take part in the interaction, regardless of whether their reply directly addressed the original thread poster or any other user participating in the thread. As mentioned above, within a given thread each user is directly connected to other users, as they participate in the same discussion. This is shown by the connections between U1, U2 and U3 in Thread 1 (T1) in Figure 2. U3 takes part in two other thread discussions (T2 and T3), hence is directly connected to other users within T2 and T3. U1 and U4

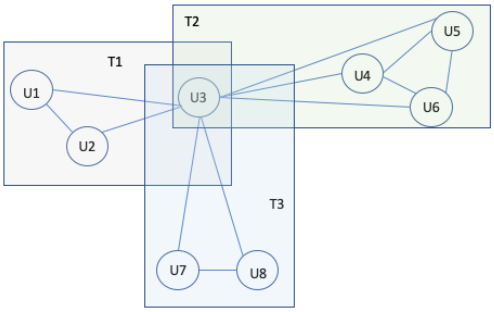

Figure 2. Network Excerpt demonstrating Users (U, Nodes), Threads (T, Boxes) and Interactions (Edges).

are two hops from each as they are both connected to U3 who participates in both T1 and T2. U1 and U4 do not participate in the same thread discussions, therefore no direct link exists between them. The more threads a given user participates in, the denser the connections.

We note that this network setup is a simplification of complex interaction scenarios, which cannot be reconstructed based on the original interaction data due to the informal nature of discussion on forums, such as which author replies to which author within the thread. It does not take into account that an author might reply to another author multiple times in the same thread, or that an author might not directly reply to another author who participates in the same discussion. Thus, the network attributes the same importance to all connections between authors participating in the same thread regardless of the number of interactions between them, resulting in some missing and some additional links. This has a direct impact on the *density* of the network as the number of relationships in the interaction network is possibly higher than the actual number. Another implication is on the analysis of *triadic closures within threads*, that is the likelihood of two nodes being connected given an existing connection between each of the nodes and a third node. Finally, the network model contains public interactions, and excludes private ones. Thus, it inherently presents a partial view of interactions taking place in the selected dark web forums.

We apply this network model to each forum resulting in six undirected networks. Following an exploratory analysis and network pre–processing, we analyse the networks, which is described in the remainder of the paper.

4.4. Network Pre–processsing

The raw network was pre–processed to produce the final network used in the analysis. First, exploratory analysis was performed to assess network scale (to ensure our sample is representative and is a suitable size for processing), and identify nodes and relationships that are not relevant in the analysis: periphery nodes that represent authors who do not actively post are discarded.

The analysis is restricted to a given time period based on post and interaction activity throughout the lifespan of the forums. For each forum we selected the time

period February 2019 — July 2019, inclusive. Thus, our analysis is based on a snapshot of forum interactions, which represents current posting activity and does not take into account all historical data present on the forums. We bound our analysis to this time period for two reasons. First, this time period is ideal of comparisons across forums due to the overlap in posting activities. Second, this time period contains the most recently collected data, which provides relevant and latest updates on the social structures and networks of the chosen forums.

Network Analysis Tools. Relevant forum data was extracted from the relational datastore and was then imported to the Neo4j graph database[3] to carry out subsequent analysis. Since data can be exported from Neo4j in multiple formats, it enables a flexible methodology to analyse the resulting social network with a plethora of graph tools. The tools we used in this study include Pajek[4], Gephi[5], NetworkX[6] and python-igraph[7].

5. Network Analysis

5.1. Network Statistics

To start our analysis, we report on basic network statistics of the largest connected component, described in Section 5.2. Specifically, we analyse *network scale* (number of nodes and edges), *density*, *average* and *maximum degrees*, *degree assortativity*, *network diameter*, and *average path lengths* as listed by Table 2. Network diameter, and average path lengths are discussed separately as part of the large-scale structural analysis in Section 5.2.

Network density, actual connections divided by potential connections, indicates the extent to which members are connected to each other by direct relations [19]. Density has been linked to a number of concepts relevant in the study of social networks. One area it has implications on is social control, that is, constraints placed on an individual's decisions through others' opinions and influence [20]. High network density within a community has been found to increase its ability to exert control on behaviour as community members are more apt to monitor and respond to such behaviour. Alternatively, low network density and weak social ties can result in the inability to exert social control [21]. Another widely researched area in the analysis of networks is the spread of information and diffusion of ideas and opinions. A group of models aimed at understanding information spreading mechanisms utilise a network-based approach studying structural characteristics of networks [22]. Literature in this space has confirmed that information spread is faster in more densely interconnected networks [23]. Since the studied networks are content-centric, that is, users connect with each other based on common interests, analysing network attributes that affect information spread is useful in providing insights on user engagement [24]. Users are more engaged when deriving value from participating in forum discussions, which is linked to the quality, novelty and relevance of information and ideas discussed. Directly related to density, we also analyse the *average degrees* of the individual networks.

Finally, we report on *assortativity* values, which indicate whether nodes tend to connect to nodes that are similar. Specifically, we are interested in *degree assortativity*, which measures similarity with respect to node degree. This measure provides insights on the pattern of connections among members of similar degrees and reveals whether they are more likely to interact. A network is assortative when high degree nodes are connected to high degree nodes and low degree nodes are connected low degree nodes [25]. Assortativity has implications on network robustness. In an assortative network, the removal of a high degree node does not affect the connections between other high degree nodes, compared to a *disassortative* network, where high degree nodes do not have dense connections with each other, and there is a higher chance of the network to become disconnected [25].

5.2. Structural Analysis

A key component of the analysis is unveiling the large-scale structure of each dark web network. By taking a look at these networks from a macro perspective, we can understand the structures in which the individual forum members are nested within.

Networks take on numerous structures. In terms of degree distribution on one end of the spectrum, *Poisson random graphs* are characterised by a homogeneous degree distribution, no hierarchical patterns and an even topology. By introducing hierarchy, and as *hubs* — nodes with a larger number of connections — appear, we move towards decentralised networks. These graphs are likely to possess the *small-world property* since hubs make it possible to connect a large number of nodes, hence average path lengths shorten. Finally, on the other end of the spectrum, centralised networks show a larger disparity between nodes, a few highly connected nodes and a large number of nodes with relative low connectivity, which can be characterised by a *power law distribution*. These networks are often called *scale free networks* [26].

As part of the structural analysis, we look at 1) *degree distributions* to highlight whether the dark web networks we examine exhibit a *power law distribution*. Revealing whether dark web networks contain hubs, how their degrees are distributed, and analysing high degree nodes versus low degree nodes is useful in understanding some key interaction mechanisms present in these forums.

In conjunction, we investigate 2) if the *small-world effect* is applicable to these forums. The presence of this phenomenon is particularly interesting in the context of online forums, as structure is one of the factors that enables information, such as new hacking techniques or crime related posts, to spread in the network, which we also discuss in the context of density in Section 5.1. This is particularly relevant when discussing criminal activities in underground forums.

Third, we identify 3) *network components*, a subset of vertices in which each node is reachable from all other nodes in the subset, and explore the level of integration (or fragmentation) in the networks. Similar to the small-world effect, the level of connectedness of these interaction

3. http://www.neo4j.org/
4. http://vlado.fmf.uni-lj.si/pub/networks/pajek/
5. https://gephi.org/
6. https://networkx.github.io/
7. https://igraph.org/python/

networks impacts the way information flows and allows the discovery of disconnected parts of the networks.

Finally, we identify subgroups and community structure within the networks via *community detection*. To detect dense sub-groups, we utilise the *Louvain algorithm*, which optimises *modularity*. Modularity measures the edge density within a community "as compared to links between communities" [27]. A sub-group is characterised by denser connections within the group compared to outside the group. Clusters are characterised by a unique set of properties. For example a subgroup that is part of a forum network and members of the subgroup may share a particular ideology that is different from the ideology of another subgroup within the same forum. Modularity is also researched in other contexts in network analysis, such as measuring network robustness. Although we do not perform this analysis in the current study, it is a particularly interesting question for dark web forums if the network is robust against interventions.

5.3. Network Centrality

Next, we measure network centrality within the largest components aimed at identifying influential members of the forums. Out of the various centrality measures we analyse *betweenness, closeness* and *eigenvector centralities*, where each measure defines importance in a different manner. While overall degree centrality counts the number of connections a node has, *eigenvector centrality* takes this a step further by attributing more importance to a node if its neighbours themselves are influential. *Betweenness centrality* helps identify so called 'bridge' nodes, where influence stems from their ability to connect sub-communities. Finally, *closeness centrality* discovers nodes that occupy a position in the network, by having a low distance to other nodes, that allows them to broadcast information [26].

We complement the quantitative findings with an analysis of the posting activities of nodes with high centrality scores. Firstly, we investigate whether there is a correlation between centrality scores and posting activity of these members, followed by observing the sub-forums they post in. Finally, we perform a qualitative analysis of the posts of these users to reveal themes that appear in them, and the roles these users might play on the forums. Since this step centres around analysing the content of posts (through a random sample for each forum), we exclude non English speaking forums.

6. Results

6.1. Network Statistics

Two of the networks we analyse, Forum B and Forum E, are relatively small compared to the medium-sized ForumsA, D and F, and the largest network, Forum C. However, Forum B presents a richer picture of interactions than Forum E, as it has a larger number of edges.

Table 2 shows that the larger the network, the smaller the density, and networks of similar sizes have similar density scores. As mentioned in Section 4, our network setup has a direct impact on density scores, however, they still provide a useful indication of how well-connected the networks are overall. Density is affected by the ease of creating connections, in this case the ability to interact with other members by joining existing discussions or creating new threads. The more threads authors post in together, the denser the network, which has implications on the ease of information spread. As mentioned in Section 5.1, ideas spread faster in more densely connected networks. Thus, a new hacking technique posted by a random user on the network has a higher chance of reaching another random member on Forum B compared to Forum C. Compared to previous studies on carding networks [28], the densities for *Forums A, C, D, F* were comparable to previous findings on carding forums. *Forums B and E* were more connected in comparison.

While members with the highest degree scores connect to a small fraction of their respective networks, the member with the highest degree score in Forum E is connected to 24% of the remaining nodes, which indicates that the user is highly active and participates in a large number of interactions compared to others. Since Forum C is so large in comparison, reaching $\sim 2\%$ of the users may outweigh the influence of the highest degree member in Forum E.

The assortativity values of Forum C, E and F are negative but not significantly lower than zero. Forum A, B and D are disassortative networks, which has implications on how easily these networks could become disconnected by removing key nodes. This also suggests that in these forums low degree nodes take part in the same discussions, defined by interests, as high degree nodes. Additionally, high degree nodes might contribute to threads related to specific subjects, and might not discuss the same topics. Overall, the interactions across nodes of varying degrees suggest cross-interactions between nodes of dissimilar connectivity.

6.2. Network Structure

6.2.1. Degree Distributions. When looking at the frequency distributions of node degrees, in Figure 3, all forums share a characteristic, that is, their degree distributions are left–skewed: the majority of nodes have low degrees, while a relatively few nodes with high degrees are on the tail of the distribution. Forum E follows a different degree distribution to the other forums, and we hypothesise that this might be due to the size of the network, which contains 372 nodes altogether.

6.2.2. Small World Effect. In the context of the small world effect usually two properties of the large scale structure of forums are analysed. First, the *clustering coefficient*, where a high clustering coefficient value indicates that many of the connections of a given node are also connected [5] [8]. Second, the *average shortest path length* shows on average how many hops it takes to get from a node to another one in the network [5] [8].

As mentioned in Section 5, as a result of the network model construction, the links within threads connecting authors are at a maximum (all authors are connected to all other authors within the same thread), which affects both density and the average clustering coefficient by increasing the value of these scores. The degree by which

	Forum A	Forum B	Forum C	Forum D	Forum E	Forum F
Author Node Count	1,552	373	16,401	1,781	22	2,887
Interaction Edge Count	15,159	3,900	624,926	19,636	57	63,688
Density	0.012	0.056	0.004	0.012	0.247	0.015
Average Degree	19.53	20.91	76.2	22.05	5.18	44.12
Maximum Degree	699	188	15,617	628	14	1,202
Assortativity	-0.18	-0.23	-0.035	-0.27	-0.08	-0.08
Network Diameter	6	4	8	6	4	6
Average Path Length	2.5	2.3	2.8	2.6	2.08	2.5
Clustering Coefficient	0.771	0.769	0.687	0.83	0.804	0.665

TABLE 2. INTERACTION NETWORK STATISTICS

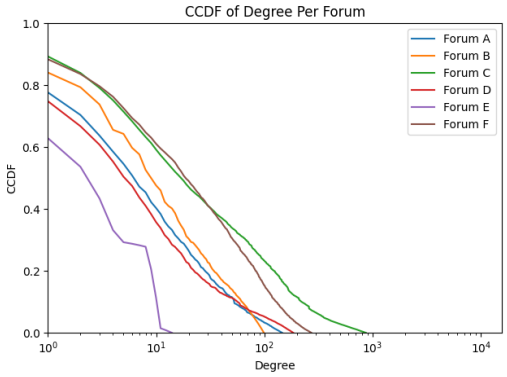

Figure 3. CCDF showing Degree Distributions on each Forum

the scores are increased depends on the interaction scenario. In some cases the difference might be negligible, while in others additional links are added that would otherwise not be part of the network. Thus, we do not heavily rely on the clustering values to decide whether the networks exhibit a small world property. However, when comparing the clustering coefficients of the interaction networks with corresponding metrics in generated *Erdős-Rényi random graphs*, based on the differences the conclusion can be drawn that the corresponding networks exhibit different properties in this sense. These differences would be present and sufficiently large even with a drop in the average clustering coefficient scores of the interaction networks. The average clustering coefficient of the generated random graphs ranges between *0.004 – 0.2*, while the coefficient values of the interaction networks range between *0.6 – 0.8*. The difference between the average clustering coefficient suggests that the small-world model does not fully explain the formation of the observed networks.

As shown in Table 2, the average shortest path length is around 2, that is, any member of a forum interacting in a given thread is two hops from another member interacting in any other thread. In other words, any two randomly selected member, *x* participating in thread *t1* and *y* participating in thread *t2*, are connected to at least one member who participates in both *t1* and *t2*. This is true for the smaller and larger networks alike. Diameter values for all networks except Forum C range between 4-6 hops and stand for the longest of the shortest paths indicating how far information needs to travel from one end to the other end of the network. Forum C has a larger network diameter with a value of 8. Similarly to network density, lower average path lengths facilitate the spread of information and ideas. This might for example take the form of user U3, shown by Figure 2, posting a citing of thread T3 discussed by U7 and U8 to a question asked by user U6 in thread T2. In the example the information shared in T3 is now shared and potentially further discussed in T2, which was made possible by node U3 and that U7 and U8 are two hops from U6.

6.2.3. Network Components. As shown in Figure 4, the *Forum E* interaction network is fragmented and consists of 47 components. Since most social networks have a largest component that accounts for the majority of nodes [26], Forum E shows a different pattern. There are four larger components, each consisting of 15-22 nodes. The largest connected component (22 nodes) accounts for $\sim$ 10% of the entire network (207 nodes). In this component there are 4 hubs, which are connected to each other through multiple links indicating frequent interactions. Their corresponding degrees are 28, 25, 23, 17, while the remaining nodes have considerably lower degrees. The authors corresponding to these hubs have a relatively higher reputation score and a larger number of posts compared to others. An analysis of thread titles shows that discussions revolve around hardware related subjects.

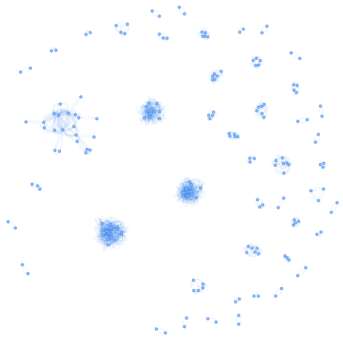

Figure 4. Forum E Connected Components.

Forum B, similarly to Forum E, is a smaller network. However, it shows a fundamentally different pattern of connections. It has a single largest component that accounts for most of the network. The largest component contains 373 out of a total of 379 nodes. The remainder of the network represents a couple of isolated discussions with 2-3 participating nodes on the periphery.

Forum A, Forum C, Forum D, and *Forum F* are larger than the previous two forums, and share the same attribute that they have a largest component that accounts for most of the network. The periphery components consist of 2-4 nodes, which represent discontinued threads. Forum F's largest component consists of 2887 nodes (out of a total

of 2912). The component sizes for Forum A, Forum C, and Forum D are as follows: 1552 out of 1576 nodes and 12 connected components, 16401 out of 16526 nodes and 58 connected components, and 1781 out of 1819 nodes and 19 components, respectively.

6.2.4. Community Detection. Using the modularity algorithm in Gephi, which finds nodes that are more densely connected together than to the rest of the nodes in the network, the aim of this analysis is to identify modules. The algorithm produces an overall modularity score, which defines the strength of the divisions in the network. A higher modularity score indicates denser connections between nodes within the modules.

To begin community detection, different configurations of the modularity algorithm were tested and each network was divided into multiple modules. The number of modules yielded for each forum ranged from nine to 18, with the exception of Forum E with two modules. The percentages of nodes within the top three modules across forums suggest that most nodes within each forum were accounted in these densely connected communities. For example, Forum A, with a total of 18 modules, accounted for 88.33% of the nodes across its top three modules. As the networks are based on interactions within threads and threads are centered around a given subject, we hypothesise that the network divisions might correspond to members participating in discussions based on their interests. This brings forth the question whether most users specialise in specific areas or if they participate in discussions on various subjects. An analysis of the modules within Forum E shines a light on this from the perspective of a very small community.

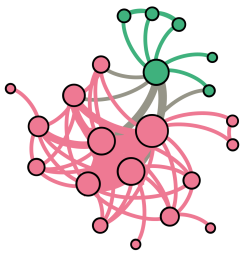

Figure 5. Forum E Communities.

The largest connected component of Forum E is a network that consists of merely 22 nodes. We have identified two modules, as shown in Figure 5. Since it is such a small network, we do not claim that it is representative of other dark web forums. However, due to their small size, we can perform content analysis on the two modules. By analysing posts, post count, and author degrees, we can see that the core of the first, larger module consists of author nodes who have posted more and who present a wider range of interests as opposed to a niche subject. These authors seem competent in multiple areas, and their distinct attribute is that they provide help, answer questions, and express opinions in their posts with occasional questions. They are characterised by a larger involvement and also shape the quality of discussions. Some posts demonstrate personal interactions with references to other forum members by their usernames. The second, smaller module on the other hand consists of a single node that represents more posting activity and connection to other more active members, while the rest of the nodes that connect to it, show a distinct pattern of posting behaviour. These members ask for help and post questions about a specific subject (torrents) and they do not take part in longer discussions. An interesting question is what would happen if we were to remove the nodes with the highest degree in these modules. In this particular forum and within the largest component when taking into account this static view of network, the result is likely to be a change in the nature of the interactions. Most posts would ask questions, and most posts expressing opinions would disappear, potentially disrupting the flow of information and knowledge.

6.3. Network Centrality

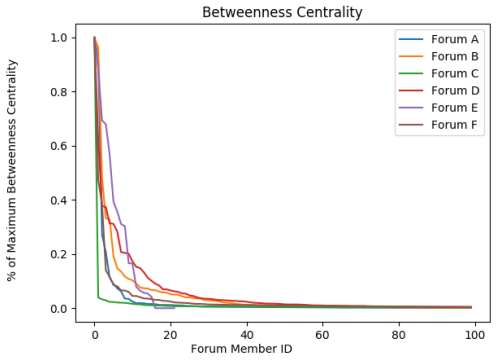

Figure 6. Top 100 Members of Betweenness Centrality

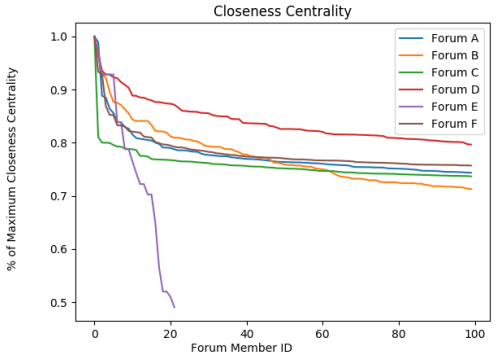

Figure 7. Top 100 Members of Closeness Centrality

1. Quantitative analysis. Across all forums, the network all-degree centralisation suggests a center existing within the largest components of all six forums to an extent. Forum E has the highest all-degree centralisation of 0.462 while Forum D has the lowest value of 0.341.

When plotting betweenness (Figure 6), closeness (Figure 7), and eigenvector centrality (Figure 8) for the highest 100 nodes in the largest component, we observe differences in the rate of change for each of these measures. For betweenness centrality, Forums A, B, D, and F all gradually decrease in centrality for later members. Forum C rapidly decreases in centrality after the first member,

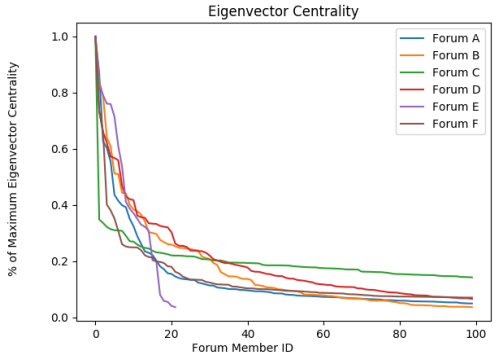

Figure 8. Top 100 Members of Eigenvector Centrality

and Forum E gradually decreases, but is considerably smaller in size than the other forums. Closeness centrality highlights a difference with Forum D, which maintains a higher proportion of centrality for other members compared to other forums. For eigenvector centrality, we find the tail of Forum C has a higher proportion of centrality compared to other forums. Also, Forums B and D decrease at a slower rate than Forums A and F, although both have similar proportions at the tail of the curve.

Across the largest components of each forum, there were overlaps in the identification of central nodes across all three centrality measures. In particular, the centrality measures agreed on at least three of the top six central members. Forum A was the only forum where the key nodes were consistent across all measures. These findings suggest most users who were highly reachable within these forums were simultaneously in control of the information flows within the largest components.

To determine if users' centrality is related to their posting activity, we performed basic *Pearson's correlation tests* between all centrality measures with posting activities for central members across all forums. Given the construction of our networks, we measured posting activities as the number of posts and number of threads created by users.

Results showed significant positive correlation (with r values greater than 0.72) for the majority of forums between *betweenness centrality* and posting activities, and between *eigenvector centrality* and posting activities. For both measures of posting activities, *closeness centrality* did not achieve significance and showed negative correlations. Within Forum E correlation between posting activities and *closeness* and *eigenvector centrality* measures failed to achieve significance. These results further highlighted different structural features and dynamics at play for central users in Forum E. It is possible that they do not simply rely on active posting to achieve their status.

2. Qualitative analysis. The second part of our centrality analysis takes a qualitative approach. First, we analyse the posting behaviour of the top nodes based on their centrality scores. This is followed by a characterisation of top nodes based on themes we identify in their posts.

Posting behaviour of top nodes. The most diverse forum in terms of posting activity is *Forum B* as all high betweenness and closeness centrality nodes post in multiple boards related to various subjects. However, the majority of their posting activity takes place within specific sub-forums. Three of the top users post in the general discussion sub-forum, while the other three users post in the 'Weapons' board. The posting behaviour of high eigenvector centrality nodes is very similar, with the exception of one member who is primarily active in the 'News' board.

In *Forum C* high betweenness and closeness centrality nodes tend to post in various sub-forums, although some members focus on a few specific boards. This might indicate some level of 'specialisation'. Four of the top members post in the board titled *'Darknet Markets'* the most, while the two remaining members post in *'Xanax'* and a board titled [Forum C]. Although *'Darknet Markets'* and *'Xanax'* are specific to a subject, they belong to the largest boards of the forum based on the number of threads within these sub-forums. Analysing the high eigenvector centrality nodes highlights that half of them are active in a marketplace board.

Top nodes within *Forum A* post mostly in a few selected sub-forums and not in a wide range of boards as seen with the previous forums. The majority of posting activity stems from the *'Beginners'* and the general discussion sub-forums. A majority of the posting activities of top nodes in *Forum D* occurred within the *'Tutorials'* sub-forum. In *Forum E* almost all posting activity takes place within the *'Hardware and Networking'* board. In *Forum F* three high betweenness centrality nodes post in a variety of sub-forums, while the remaining three focus on one or two. All six show the most activity in the sub-forum *'General Discussions'* and some activity in the marketplace discussions board.

Characterisation of top nodes. Investigating the content of discussions posted by high centrality score members revealed information about these users, and distinct user profiles emerged. These profiles allow us to hypothesise the role these users might play in the forums.

High centrality score members on *Forum A* discuss a wide variety of topics including marketplaces, cryptocurrencies, gambling, drugs, and sexual relationships. They are also heavily involved in discussions related to staying anonymous on the dark web and predominantly share advice. We also identified one of these users to be a moderator setting the rules of the forum and maintaining control through frequent posting.

Users corresponding to top nodes in *Forum C* participate in discussions frequently. Their topics revolve around general hacking and drug markets, where they are often involved as buyers or vendors. Many of these members participate in giveaways, which leads to activity in many threads. A majority of members in the group discuss operational security to avoid having their activities alert law enforcement. Topics of these posts include technical details of encryption, escrow services and scams, and using different types of cryptocurrencies to hide activities. Some of the posts promote various markets when other markets have been under DDoS attacks. A minority of members carry out moderator activities.

Posts of high centrality score members on *Forum E* show that this forum has different characteristics compared to the others. Firstly, discussions of high centrality users do not tend to revolve around topics of cybercrime. One particular member with high centrality scores sim-

ply engages in a large number of interactions by asking questions related to computing in general and specifically about hardware. The interactions also involved building rapport with the community, responding to posts that were replies to their questions, engaging in arguments on the subject, and sharing information with the community. The other group of high centrality score users were on the other end of the same interaction through sharing their knowledge and providing advice on hardware related questions. One particular member mentioned that they were new to the forum. Thus, members in this forum can become high centrality score users simply by taking part in a large number of interactions and asking questions to gain knowledge of a particular area.

Forum F presents a more colourful picture of the top nodes. One role that emerges is the user who provides detailed technical advice to others in a large number of topics related to hacking, general security, marketplaces. Their posts suggest deep technical knowledge and forum engagement. Another role top nodes play is characterised by active participation in different threads by asking questions and providing advice to high level questions while not specialising in a particular technical area or activity, such as being a vendor. These users directly refer to other members, have conversations with them, and greet them to the forum in their posts. These users seem to play a community building role, and they also express their opinions about the general directions of the forum, politics and freedom. Some high centrality users seem to be well-versed in subjects related to the dark web, including marketplaces. They mostly provide advice in these areas and might facilitate deals. Another user profile we discovered is that of a user whose main interests are marketplace related. These users share detailed information on Tor anonymity, advice on getting started on marketplaces, and different cryptocurrencies. The counterpart of this role is the user who asks for feedback from the community about technical solutions they proposed, and questions about marketplaces. Finally, we identified some users to be moderators on this forum as well.

7. Discussion

Posting activities of central nodes. Individuals with a significant interest in hacking participate in multiple online communities in order to gather more information and increase the density of their social networks [29]. It is not within the scope of this work to investigate the movement of users between forums. However, we found evidence of central nodes participating in multiple sub-forums within a single forum. For example, while central members in *Forum D* posted mostly in the 'Tutorials' sub-forum, top nodes in Forum F posted across different boards showing a wide array of interests. These differences also suggest that highly connected nodes play different roles in the forums we analyse, underlined by the results of our qualitative analysis in Section 6.3.

Given the sub-forums where central nodes posted, it is possible that some of these forums, such as Forums A, D and F are more welcoming of new users. This is reflected in the sparse connections across the networks of these forums. In addition, the network assortativity of Forums D and F indicated interactions between high degree nodes and low degree nodes, which suggest a welcoming community, and the possibility of information and resource flow between dissimilar users, such as between more experienced and newer users. Interactions among new members and experienced ones can shape opinions and ideas and potentially influence the behaviour of new members [9]. On the other hand, Forum B and C tended to focus more on marketplace-related discussions, suggesting that interacting with new members is less of a priority. At the other end of the spectrum is Forum E where it appears to be a tight-knitted community with specific interests.

A similarity we found on all forums is that central players are not uniform, and multiple user profiles exist, ranging from moderators, technical gurus, community building members to marketplace actors.

Structural attributes and implications on disruption and information spread. The effectiveness of disruption strategies is known to depend on both network topology and network resilience [30]. Current findings on network structures and interactions, as well as the activities of central nodes, provide some food for thought on existing approaches to monitor and disrupt these dark web forums. The removal of key players is a common approach to disrupt online marketplaces [8], [31]. A factor that affects the effectiveness of this strategy is network centrality. Previous studies have shown that the removal of high-value targets in decentralized organisations, which dark networks often are, does not always shut them down but sometimes drives them to become more decentralised, making them even harder to disrupt. It is suggested that criminal networks might even become 'stronger', after targeted attacks [30]. Another class of disruption strategies aim to disrupt marketplace dynamics via information. An example is known as 'lemonising the market' [31], [32]. This strategy targets online marketplaces with the purpose of creating distrust between actors since a lemon market is one with uncertainty over product quality [33]. 'Lemonising the market' is dependent on the efficiency of information spread within the network, which in turn is affected by network density; a denser network increases enforceable trust (i.e. compliance to group expectations by users over desires-driven behaviours) [21].

Our findings suggest that the first disruption approach has the highest potential impact on forums comparable to Forum E, while the second disruption approach is more suitable for forums with network structures comparable to the remaining five forums. The fragmented network structure of Forum E points to the concentration of information within each large component and thus the potential lack of redundancies in information or member roles across the forum [28]. Thus, removing key actors would highly disrupt information flow within such networks.

The network structures of the remaining forums suggest their vulnerability towards disruption via information. The inclusive large components and low density of these forums increase their resilience against the first type of strategy due to redundancies in information and roles of members [28]. More specifically, the flow of information tends to occur within sub-forums focusing on tutorials and general discussions. In addition, the high average degrees and short average path lengths suggests that information travel far and fast within the largest components [26]. In this sense, it may be more cost-effective to disrupt via

information, such as the spread of false information to increase offenders' efforts or posting real information on law enforcement to increase perceived risks [31].

8. Conclusion

In this study we investigated the differences between six dark web forums using social network analysis metrics and analysis methods. We observe the large scale structure and influential nodes within these networks. This work can be taken forward in a number of directions. The static network analysis could be complemented by analysing an evolving network of interactions and network growth patterns. Additionally, the attributes of the various node types allow a fine-grained analysis to be done based on the values of these attributes. For example, an analysis of a sub-network could be carried out based on dividing the network to discussion subjects.

References

[1] T. J. Holt and E. Lampke, "Exploring stolen data markets online: products and market forces," *Criminal Justice Studies*, vol. 23, no. 1, pp. 33–50, 2010. [Online]. Available: https://doi.org/10.1080/14786011003634415

[2] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker, "An analysis of underground forums," in *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, ser. IMC '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 71–80. [Online]. Available: https://doi.org/10.1145/2068816.2068824

[3] M. Yip, N. Shadbolt, T. Tiropanis, and C. Webber, "The digital underground economy: a social network approach to understanding cybercrime," in *Digital Futures*, 10 2012.

[4] S. Pastrana, A. Hutchings, A. Caines, and P. Buttery, "Characterizing eve: Analysing cybercrime actors in a large underground forum," in *International Symposium on Research in Attacks, Intrusions, and Defenses*. Springer, 2018, pp. 207–227.

[5] J. Xu and H. Chen, "The topology of dark networks," *Communications of the ACM*, vol. 51, no. 10, pp. 58–65, 2008.

[6] H. Chen, *Dark Web. Exploring and Data Mining the Dark Side of the Web*. Springer-Verlag New York, 2012.

[7] R. Frank, B. Westlake, and M. Bouchard, "The structure and content of online child exploitation networks," in *ACM SIGKDD Workshop on Intelligence and Security Informatics*. ACM, 2010, p. 3.

[8] M. Yip, N. Shadbolt, and C. Webber, "Structural analysis of online criminal social networks," in *2012 IEEE International Conference on Intelligence and Security Informatics*. IEEE, 2012, pp. 60–65.

[9] T. Holt, D. Strumsky, O. Smirnova, and M. Kilger, "Examining the social networks of malware writers and hackers," *International Journal of Cyber Criminology*, vol. 6, 01 2012.

[10] M. Macdonald and R. Frank, "The network structure of malware development, deployment and distribution," *Global Crime*, vol. 18, no. 1, pp. 49–69, 2017.

[11] S. Samtani, "Using social network analysis to identify key hackers for keylogging tools in hacker forums," 09 2016, pp. 319–321.

[12] V. Griffith, Y. Xu, and C. Ratti, "Graph theoretic properties of the darkweb," 2017.

[13] B. Monk, J. Mitchell, R. Frank, and G. Davies, "Uncovering tor: An examination of the network structure," *Security and Communication Networks*, vol. 2018, pp. 1–12, 05 2018.

[14] M. De Domenico and A. Arenas, "Modeling structure and resilience of the dark network," *Physical Review E*, vol. 95, no. 2, Feb 2017. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.95.022313

[15] E. Phillips, J. Nurse, M. Goldsmith, and S. Creese, "Extracting social structure from darkweb forums," 11 2015.

[16] B. da Cunha, P. MacCarron, P. Passold, J. W. dos Santos Jr., K. Oliveira, and J. Gleeson, "Assessing police topological efficiency in a major sting operation on the dark web," *Scientific Reports*, vol. 10, no. 73, 2020.

[17] M. Zamani, F. Rabbani, A. Horicsányi, A. Zafeiris, and T. Vicsek, "Differences in structure and dynamics of networks retrieved from dark and public web forums," *Physica A: Statistical Mechanics and its Applications*, vol. 525, p. 326–336, Jul 2019. [Online]. Available: http://dx.doi.org/10.1016/j.physa.2019.03.048

[18] S. Pastrana, D. R. Thomas, A. Hutchings, and R. Clayton, "CrimeBB: Enabling cybercrime research on underground forums at scale," in *Proceedings of the 2018 World Wide Web Conference*, ser. WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, p. 1845–1854. [Online]. Available: https://doi.org/10.1145/3178876.3186178

[19] R. J. Sampson, "Neighborhood and crime: The structural determinants of personal victimization," *Journal of Research in Crime and Delinquency*, vol. 22, no. 1, pp. 7–40, 1985.

[20] B. Janky and K. Takács, "Social control , participation in collective action and network stability," 2002.

[21] K. A. Frank and J. Y. Yasumoto, "Linking action to social structure within a system: Social capital within and between subgroups," *American journal of sociology*, vol. 104, no. 3, pp. 642–686, 1998.

[22] J. L. Iribarren and E. Moro, "Affinity paths and information diffusion in social networks," *Social Networks*, vol. 33, no. 2, p. 134–142, May 2011. [Online]. Available: http://dx.doi.org/10.1016/j.socnet.2010.11.003

[23] B. Liu, R. Madhavan, and D. Sudharshan, "Diffunet: The impact of network structure on diffusion of innovation," *European Journal of Innovation Management*, vol. 8, pp. 240–262, 06 2005.

[24] V. Chaoji, S. Ranu, R. Rastogi, and R. Bhatt, "Recommendations to boost content spread in social networks," in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 529–538. [Online]. Available: https://doi.org/10.1145/2187836.2187908

[25] R. Noldus and P. Van Mieghem, "Assortativity in complex networks," *Journal of Complex Networks*, vol. 3, no. 4, pp. 507–542, Dec 2015.

[26] M. Newman, *Networks*, 1st ed. Oxford University Press, 2010.

[27] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, Oct 2008. [Online]. Available: http://dx.doi.org/10.1088/1742-5468/2008/10/P10008

[28] T. J. Holt, O. Smirnova, and Y.-T. Chua, *Data thieves in action: Examining the international market for stolen personal information*. Springer, 2016.

[29] T. J. Holt, *Lone Hacks or Group Cracks: Examining the Social Organization of Computer Hackers. In F. Schmalleger M. Pittaro (Eds.), Crimes of the Internet*. Upper Saddle River, NJ: Pearson Prentice Hall, 2009.

[30] P. Duijn, V. Kashirin, and P. Sloot, "The relative ineffectiveness of criminal network disruption," *European Journal of Innovation Management*, p. 4238, 04 2015.

[31] A. Hutchings and T. J. Holt, "The online stolen data market: disruption and intervention approaches," *Global Crime*, vol. 18, no. 1, pp. 11–30, 2017.

[32] M. K. Hoe, S. C. and A. Bensoussan, *A Game Theoretical Analysis of Lemonizing Cybercriminal Black Markets. In Decision and Game Theory for Security, edited by J. Grossklags and J. Walrand*. Springer: Berlin, 2012.

[33] G. A. Akerlof, "The market for 'lemons': Quality uncertainty and the market mechanism," *The Quarterly Journal of Economics*, vol. 84, no. 3, pp. 488–500, 1980.