Decision Support Systems 184 (2024) 114271

Contents lists available at ScienceDirect

Decision Support Systems

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

Understand your shady neighborhood: An approach for detecting and investigating hacker communities

Dalyapraz Manatova [a,*], Charles DeVries [b], Sagar Samtani [c]

[a] Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA
[b] Carnegie Mellon University, PA, USA
[c] Kelley School of Business, Indiana University, Bloomington, IN, USA

ARTICLE INFO

Keywords:
Hacker forum
Community detection
Cyber threat intelligence

ABSTRACT

Cyber threat intelligence (CTI) researchers strive to uncover collaborations and emerging techniques within hacker networks. This study proposes an empirical approach to detect communities within hacker forums for CTI purposes. Eighteen algorithms are systematically evaluated, including state-of-the-art and benchmark methods for identifying overlapping and disjoint groups. Using discussions from five prominent English hacker forums, a comparative analysis examines the influence of the algorithms' theoretical foundations on community detection. Since ground truths are unattainable for such networks, the study utilizes a multi-metric strategy, incorporating modularity, coverage, performance, and a newly introduced quality measure, Triplet Hub Potential, which quantifies the presence of influential hubs. The findings reveal that while modularity optimization algorithms such as Leiden and Louvain deliver consistent results, neighbor-based expanding techniques tend to provide superior performance. In particular, the Expansion algorithm stood out by uncovering granular hierarchical community structures. The ability to investigate these intimacies is helpful for CTI researchers. Ultimately, we suggest an approach to investigate hacker forums using community detection methods and encourage the future development of algorithms tailored to expose nuances within hacker networks.

1. Introduction

Cybercrime has become a significant global economic concern, with an annual cost of approximately $600 billion, equivalent to nearly 1% of the global GDP [1]. As the cyber threat landscape expands, businesses and organizations are increasingly implementing cyber threat intelligence (CTI) frameworks to analyze cybercriminal activities and security risks. One of the approaches is to monitor the criminal ecosystem, which has evolved from knowledge contribution on hacker forums to marketplaces structured as forums, where cybercriminals profit from selling information, databases, stolen credit cards, illicit materials, and software exploits.

Hacker forums, while structurally similar to traditional discussion forums, fulfill a specialized role within the cybercriminal community. These platforms are less about fostering open-source knowledge and more about facilitating the trade of tools and techniques, including hacking tools, exploits, and leaked data. Users engage not only in the development of personal brands, but also in the establishment of professional identities for these brands, striking a balance between anonymity and recognizable branding for profit [2]. Despite the forums' market-driven nature, where "hubs" are competitors rather than collaborators, a complex social process is at play. Yip et al. revealed that these forums act as socioeconomic mechanisms, allowing social capital gains critical to economic success [3]. Yip concludes that online criminal networks possess resilience against targeted attacks, as they are structured for economic rather than social purposes. Although forums may lack an explicit infrastructure for transactions, they serve as arenas to build trust and networks of potential profit-making collaboration [4]. Furthermore, instead of focusing exclusively on the most prominent nodes in the network, more closely analyzing the communities that form can reveal much about the nature of these collaborations [5].

The cybercrime landscape has evolved from the stereotype of "lone wolf" to sophisticated organized subcultures with diverse skill sets and low-risk enterprise [6]. These communities demonstrate the industrialization of cybercrime, capable of producing and distributing malicious products on a large scale. The concept of "illicit infrastructure", as discussed by Collier et al. underscores this transformation, showing how it has revolutionized work, experiences, practices and criminal activities in these communities. This illicit infrastructure streamlines processes and supports the construction of larger-scale, higher-level

[*] Corresponding author.
E-mail addresses: dmanato@iu.edu (D. Manatova), csdevrie@iu.edu (C. DeVries), ssamtani@iu.edu (S. Samtani).

https://doi.org/10.1016/j.dss.2024.114271
Received 16 May 2023; Received in revised form 22 May 2024; Accepted 17 June 2024
Available online 21 June 2024

malicious systems, such as turnkey hacking solutions or ransomware-as-a-service [7]. However, this infrastructure is not visible on the forums. Cybercriminals, while engaging in various arenas of action for their illicit businesses, maintain stringent operational security. They manage their online identities cautiously, minimizing attention to themselves, and avoiding leaving traces across multiple online personas. However, forums do provide a place for people to build recognizable brands through nicknames or aliases [8]. Although the community aspect of user interactions may not be immediately apparent, specialized algorithms have the capability to uncover the underlying structures they are designed to detect.

Although there are various approaches to identifying prominent hackers in such forums, the detection and examination of communities remain relatively underexplored in this domain [9]. When investigating the communities formed within forums, it is crucial to consider various factors that shape their dynamics. These include the forum's culture, thematic focus, and moderation practices, along with the definition of *community* and its boundaries. These elements not only govern interactions within these digital spaces, but also play a pivotal role in shaping the analysis conducted within the realm of CTI. To capture the formation of potential collaboration, we can assume that hacking cooperation is heavily based on strong social bonds, either from repeated cooperation or developed trust. Online social networks and forums serve as platforms where individuals seeking opportunities can initiate and develop collaborations, which often extend beyond the confines of the forum [10].

Despite the need to detect communities, the detection problem is not trivial or generalizable for all networks. Naturally, social interactions differ greatly. Community detection requires a combination of mathematical, computational, and data-analytical skills. One of its key challenges is the absence of a single definition of a community, which leads to different algorithms that identify distinct communities even on the same dataset. Additionally, the quality of detection depends on the parameters used and the type of network, making comparisons of the results of the cross-study difficult [11]. This study was driven by a set of fundamental questions:

Q1: How do actors in hacker forums organize into sub-communities?

Q2: What practical methods can be employed to detect these communities within hacker forums, specifically for Cyber Threat Intelligence (CTI) purposes?

Q3: How does the choice of the detection method influence the definition and characterization of the detected communities?

We hypothesize that algorithms from the same type of approach will tend to find similar partitions if the approach is robust to less evident communities. This means that differences in approaches create different partitions.

1.1. Theoretical and practical contributions

One of the major research contributions of this study is the development of an empirical approach tailored for the analysis of communities within hacker forums for CTI purposes. In the formulation of the approach, we carefully selected 18 community identification algorithms. These algorithms include both state-of-the-art and benchmark methodologies that can identify both overlapping and disjoint groups. The algorithms were systematically classified according to their primary operating strategies. Our research involved a comprehensive comparative analysis of these algorithms using discussions extracted from five prominent hacker forums. These forums have gained recognition as essential hubs for the exchange of exploits, tutorials, tools, and information within the CTI domain. Despite their shared reputation as the go-to platform, each forum has a unique nature and social structure.

Given the inherent challenge of lacking a definitive ground truth to evaluate community detection within hacker forums, we suggest a multimetric evaluation approach. Our approach incorporated four metrics, each carefully selected to complement the limitations of the others, providing a more comprehensive evaluation of algorithm performance. In addition to established metrics, we introduced a quality metric, *Triplet Hub Potential* (T-Hub Potential), which assesses the potential presence of hubs in a scale-free nature.

Inter-algorithm comparisons revealed instances of convergence in community identifications, informing us about the strengths and weaknesses of each method. Our study concludes that employing a combination of metrics provides deeper insight into community detection than relying solely on modularity. For instance, *Performance* measure favors methods based on clustering coefficients, which tend to detect smaller, tightly knit groups. Furthermore, the detected communities are significantly influenced by the algorithm's theoretical approach. We suggest utilizing multiple algorithms but from different methodology categories, along with a multimetric evaluation strategy.

The remainder of this paper is organized as follows. Sections 2.1 and 2.2 present an overview of existing studies and theoretical premises on hacker forums and community detection techniques that are known to be utilized on real social networks. Section 3 describes the research design, testbed and algorithmic experimental setup and evaluation methods. Section 4 presents the results. Section 5 discusses the research findings, implications of the derived approach, future ideas, and concluding remarks.

2. Literature review and methodological background

2.1. Cyber Threat Intelligence (CTI) from hacker forums

In response to the growing cyber threat landscape, businesses and organizations are increasingly adopting cyber threat intelligence (CTI) frameworks to manage and analyze information related to cybercriminals and emerging security risks [19]. CTI aims to improve cybersecurity decision making by identifying new threats and key threat actors [20]. To accomplish this, CTI researchers use various tools, including open source intelligence (OSINT), to collect information on cyber threats, such as the tactics, techniques, and procedures employed by cybercriminals, as well as the impact of these threats on both organizations and individuals. OSINT serves as a valuable resource for monitoring and evaluating evolving cyber threats, such as emerging malware strains, phishing campaigns, and data breaches. It involves analyzing and evaluating information derived from underground forums, tracking the progression of hacking expertise, gauging contributions to discussions, and evaluating social trust and participation in specific incidents. The academic literature on CTI provides essential information on strategies to identify and investigate cyber threats through OSINT. These strategies include computational methods to identify malicious tools, data breaches, key threat actors, and emerging trends [13,21–26].

The advancement of CTI has been a research topic for many years. Researchers have studied data collection techniques and effective and safe ways to identify, collect, store, parse, and monitor hacker forums [20,27]. In recent years, researchers have continued to advance CTI techniques by improving methods to identify dense discussions on potential cyber threats, such as emerging hacking tools, zero-day vulnerabilities, and exploits [13,28,29]. For example, Grisham et al. focused on the extraction of mobile malware from hacker forums [22]. Other work has focused on extracting and classifying or clustering all exploits from forums using traditional machine learning approaches [13,26] and deep learning approaches [30].

However, the identification of key hackers or actors involved in cyber threats has also been a key focus area [21–25]. Recent studies have prioritized hackers according to their influence based on their contribution to threat content [18,25] and social activity [14,16,25]. For example, Biswas et al. have performed a holistic analysis of cybercriminals online, developing a mapping of their expertise, contribution to cyber threats, sentiment of messages, and lifespan, to rank key members of hacker forums [31]. Similarly, Otto et al. compared unsupervised

Table 1
Summary of recent studies focusing on communities in online hacker forums.

Date	Author(s)	Dataset	Method(s)	Objective(s)
2021	Pourhabibi et al. [12]	TN	Random walk with Surprise optimization	Identification of Communities in MultiGraphs in the context of Terrorist Network
2020	Tachaiya et al. [13]	HF	Sparse Matrix Regression, K-means	Content Clustering
2020	Pete et al. [14]	HF	Louvain	Key Members Detection
2020	Sarkar et al. [15]	HF	Louvain	Threat and Incident Prediction
2019	Huang et al. [16]	HF	Louvain	Key Members Detection
2018	Marin et al. [17]	DNM	Louvain	Community Detection
2016	Huang et al. [18]	HF	Louvain	Key Members Detection

TN = Terrorist Networks; HF = Hacker Forum(s); DNM = Dark Net Marketplace(s).

graph embedding methods to track the evolution of hackers in forums based on their expertise [32].

Identifying key individuals on hacker forums is a crucial task for both CTI and legal prosecution. However, the detection of communities within these forums has also gained significance. Researchers have increasingly directed their attention to hacking communities as a means of identifying influential members within these forums [14–18,25]. For example, Marin et al. conducted a study on DarkNet Marketplaces (DNMs) to pinpoint vendor communities engaged in selling exploits and malware. They developed a method that uses machine learning techniques and social network connections [17]. In 2021, Pourhabibi et al. introduced a community detection method (DarkNetExplorer) customized for terrorist networks with various types of connections. This method optimizes a "surprise function" in multigraphs, a statistical measure that evaluates the quality of the partition of a network into communities [12]. In our study we test one of the algorithms that uses surprise function as an optimization. To give an understanding of the research efforts devoted to community detection within hacker and other underground forums, we have summarized the key literature in Table 1. This summary encapsulates the critical aspects of these studies, including the datasets used, the methods used for community detection, and the primary objectives pursued by each study.

2.2. Community detection techniques and quality assessment

Community detection is the process of identifying clusters within a network, where nodes with stronger ties are grouped together [33]. Originally derived from information and communication theories, the concept of a community is applied to various fields such as social sciences [34], biology [35], computer science [36], and criminology [37]. The challenge in identifying these communities comes from the varied structures of networks and the algorithm-dependent definitions of what constitutes a community. Algorithms in this field are largely influenced by their foundational assumptions and are generally divided based on whether they view communities as non-overlapping distinct groups or as entities that can share nodes [11]. Approaches vary: some methods focus on optimizing quality functions to gauge and enhance the detected communities' "goodness" [38,39], while others replicate real-world network dynamics, using either divisive techniques to split networks into smaller groups or agglomerative strategies to merge individual nodes into larger ones.

Techniques such as greedy modularity optimization methods prioritize a modularity score to define the community structure [40–43]. On the other hand, stochastic algorithms introduce an element of randomness into the detection process operating on the assumption that the nature of random walks taken through the network will naturally converge on the true distribution of tightly knit groups within the network [44,45]. Other algorithms operate on the assumption that communities form cliques, tightly knit groups where every node is connected to every other node within the group, and aim to maximize the clustering coefficient [39,46]. Furthermore, there is a growing body of work exploring the enhancement of community detection by incorporating additional attributed information about the nodes and the connections between them. These advanced methods can potentially produce more specified community detection based on attributes, but require more detailed data on the components of the network or nodes, which are often nearly impossible to gather from anonymous and underground online communities, or require detailed analysis of the content posted on the forum [12,47].

In the subsequent Section 3.3, we categorize the algorithms chosen based on the algorithmic approach. We do not include in the literature and the scope of this study deep learning (DL) based methods, as DL-based algorithms might bring several limitations for the CTI and OSINT tasks, such as examining the formation of collaboration between participants of cybercriminal communities.

The reservation towards integrating DL models into our analysis of community detection hinges not only on interpretability and scalability, but also on the inherent nature of DL methodologies. Unlike traditional algorithms that often optimize a clearly defined objective function or operate on the theoretical grounding behind it, DL models derive their partitions from layered learning processes, either from embeddings of nodal attributes or/and network structures. This makes DL methods ambiguous for the categorization of the approach and, hence, the comparison with other approaches and measures of community quality. Furthermore, since some DL-based methods require training datasets, the absence of ground-truth labels complicates the use of such algorithms, but also the evaluation and comparison of performances [48]. Scalability also poses a practical challenge as DL-based models require considerable computational resources, increasing with network size, which can be impractical for CTI purposes.

However, the potential of DL algorithms to identify nuanced community structures in cybercriminal contexts warrants future investigation, where we can explore these models in more focused studies, where their effectiveness can be evaluated on a smaller scale with concrete test cases. However, the objective of the study is not only to detect communities, but also to understand the underlying reasons for their formation.

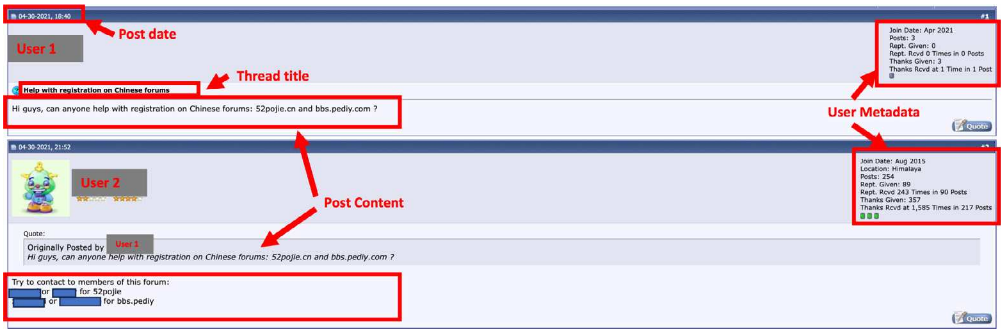
3. Research design and testbed

Our research design comprises four main components: data collection, graph construction, community detection, and a closer look at the detected communities. Each process is described in the subsequent sections below (Sections 3.1, 3.2, 3.3 and 3.4).

3.1. Data collection

We identified five English language hacker forums for collection and analysis based on the existing literature in the field [19,26,32,49–51]. These forums were chosen for their prominence as popular destinations for hackers seeking assistance and their tendency to share exploits, source codes, and tools frequently [19]. These forums, known for their varied structures and user densities, fit the objective of this study of investigating the communal structure in various cases.

We collected forum data using a custom crawler with pre-defined credentials to extract threads, posts, and metadata such as timestamps and user details. This established scraping method, reflected



04-30-2021, 18:40

User 1

Help with registration on Chinese forums

Hi guys, can anyone help with registration on Chinese forums: 52pojie.cn and bbs.pediy.com ?

Join Date: Apr 2021
Posts: 3
Rept. Given: 0
Rept. Rcvd 0 Times in 0 Posts
Thanks Given: 3
Thanks Rcvd at 1 Time in 1 Post

04-30-2021, 21:52

User 2

Join Date: Aug 2015
Location: Himalaya
Posts: 254
Rept. Given: 89
Rept. Rcvd 243 Times in 90 Posts
Thanks Given: 357
Thanks Rcvd at 1,585 Times in 217 Posts

Quote:
Originally Posted by User 1
Hi guys, can anyone help with registration on Chinese forums: 52pojie.cn and bbs.pediy.com ?

Try to contact to members of this forum:
[redacted] or [redacted] for 52pojie
[redacted] or [redacted] for bbs.pediy

Fig. 1. Example post with a reply on *exetools*, traditional hacker forum. Personal information is masked.

in prior cybercrime-related research [4,20,52,53], efficiently captures web content for analysis while maintaining user privacy, as shown in Fig. 1.

Selected forums feature threaded discussions where users can start topics and engage in unlimited replies, often leading to tangential debates. Interactions range from quick exchanges to extended dialogues with lulls. Table 2 shows the activity level of each forum, with thousands of threads and active users. Although they are English-based and long-established, forum sizes differ.

Exetools is a forum self-described as primary discussions on reverse engineering, software cracking, and computer security. It serves as a platform where experts and enthusiasts exchange knowledge, techniques, tools, and code samples in these fields, covering topics like reverse engineering, debugging, malware analysis, software security, and hacking. Note that the forum is currently closed, meaning that registration requires the administrator's approval; however, at the time of data collection, registration did not require any form of vouching.

Cardingteam, on the other hand, is known for hosting discussions related to illegal activities, particularly credit card fraud, identity theft, and other forms of cybercrimes. Participants in this forum share information on methods of committing credit card fraud, fraudulent transactions, and trading stolen credit card data and personal information. The forum also investigates hacking, money laundering, and various online scams.

Cipher focuses on cryptography, encryption, and cybersecurity discussions. Users engage in discussions about secure communication methods and seek advice on encryption-related challenges, while staying up-to-date on cybersecurity developments.

Go4Expert is an online community primarily centered around programming, web development, and technology. Members discuss programming languages, software and web development technologies, and various technical challenges. It provides a platform for programmers, developers, and tech enthusiasts to seek assistance, share knowledge, and collaborate on solving coding and development issues. However, it is not structured as a question-and-answer platform, such as StackOverflow, but a natural discussion of issues driven in a threaded manner.

Antionline is another online community that emphasizes discussions on computer security, hacking, and technology. Users engage in conversations related to hacking, network security, vulnerabilities, and programming.

It is important to note that while some discussions in these forums are centered on knowledge sharing and expertise, others may involve implicit transactions among individuals interested in specific services or tasks related to the expertise being discussed.

3.2. Graph construction

In our graph construction, we use an approach to model *implicit* social connections among users based on their contributions to thread topics initiated by other users. Specifically, we consider all users who have posted in the same thread initiated by a user to be socially linked to that user and to each other. This approach has been used in previous studies of hacker forums [14,25], and helps to capture the nature of social interactions and information flow.

Table 2
Data collection from hacker forums.

Forum	Earliest Post	Post count	Author count	Thread count
antionline	2001–07–30	466,268	13,287	58,727
exetools	2002–01–16	32,776	909	3,303
cardingteam	N/A	2,765	690	1,222
cipher	2015–05–25	42,870	3,564	3,820
go4expert	2004–07–15	76,648	14,953	19,748

The graph for each forum as a weighted directed monopartite graph $G = (V, E, \omega)$, where V is a set of nodes, (i.e., users), $E \subseteq \{(u, v) \mid (u, v) \in V^2, u \neq v\}$ is a set of directed edges from u to v, and $\omega : E \rightarrow \mathbb{R}$ is a function mapping every edge between nodes to its weight value w, defined by $\omega(u, v) = w$, where $w > 0$ if $(u, v) \in E$ and u replied with frequency w in the *thread(s)* where v participated. When there is no observed participation of u in the same discussions where user v participated, then $(u, v) \notin E$ and we set $\omega(u, v) = 0$.

Table 3 summarizes the resulting graph representations of the five selected forums. Nodes without links or degrees less than two are excluded from the analysis as we assume they do not contribute to the communal structure. It is important to note that the structural characteristics of these forum graphs exhibit variations, as highlighted by their degrees, density, average clustering coefficients, strongly connected components (SCCs) and average shortest path, all of which are summarized in Table 3.

The *average clustering coefficient* is a metric that reveals the extent to which the network's users are interconnected, essentially indicating the presence of full cliques within the network. In other words, it measures how tightly the users are connected to each other and all forums show relatively high average clustering coefficients, suggesting that users tend to form tightly-knit groups. Graph *density* represents the ratio of actual connections to all possible connections between nodes. A higher density value implies more connections within the network, signifying that users are actively engaged in the forum's activities. Note that "exetools" exhibits the highest network density, indicating strong user interactions in contrast to "go4expert", which is a sparse network. A *strongly connected component (SCC)* is a subgraph in which every pair of nodes is connected by a directed path, meaning that there is a way to connect from one node to any other within the SCC, and vice versa. In the dataset, although some forums have several SCCs, the largest SCC spans most of the network, meaning that most users are involved in the forum's main agenda and do not tend to break into evident subgroups.

3.3. Community detection algorithms

In the subsequent phase of our study, we begin to select algorithms designed to detect communities. However, prior to introducing these

Table 3
Descriptive statistics of the graph representations for each forum.

	antionline	exetools	cardingteam	cipher	go4expert
Number of nodes	11,459	886	452	3364	8774
Number of edges	4,9 M	138,590	4530	317,388	246,466
Sum of weighted edges	27 M	376,003	9516	1,25 M	622,388
Average degree	858	313	20	189	56
Max degree	10,320	1244	258	3874	12,264
Average Clustering Coef	0.91	0.71	0.85	0.84	0.86
Density	0.04	0.18	0.02	0.03	0.0032
Components (SCC)	6	1	2	1	18
Size of the largest SCC	11,444	886	449	3364	8733
Average Shortest Path (on largest SCC)	2.45	1.88	2.93	2.18	2.51

SSC = Strongly Connected Component.

Table 4
Overview of assumptions and types of algorithms used for comparison.

Algorithm (Year)	Approach based on	U	D	W	D/O	Complexity
Infomap (2008) [45]	Random Walks and Minimization of the MAP Function	Y	Y	Y	D	$O(n \log n)$
Walktrap (2005) [44]	Random Walk Probabilities	Y	N	N	D	$O(n^2 \log n)$
CPM (2011) [54]	Rewarding of Intra-community Edges and Penalize Missing	Y	N	Y	D	$O(n^q)$
Surprise (2015) [55]	Surprise Optimization	Y	Y	Y	D	$O(nm)$
SCD (2014) [39]	Maximization of Weighted Community Clustering	Y	N	Y	D	$O(km)$
RB Pots (2006) [56]	Spectral Modularity Optimization	Y	Y	Y	D	$O(n \log n)$
Louvain (2008) [40]	Hierarchical Modularity Optimization	Y	N	Y	D	$O(m)$
Leiden (mod) (2019) [42]	Agglomerative Modularity Maximization	Y	N	Y	D	$O(n \log n)$
LE (2006) [57]	Eigenvectors and Eigenvalues of the Modularity Matrix	Y	N	N	D	$O(n^3)$
EdMot (2019) [58]	Edge Enhancement Rewired to Form Cliques and Partitioning based on Hierarchical Clustering	Y	N	Y	D	$O(m^{3/2} + n \log n)$
LP (2011) [59]	Propagation of Labels through Neighbors	Y	N	N	D	$O(n + m)$
SLPA (2011) [60]	Propagation of Multi-Labels through Neighbors	Y	N	N	O	$O(in)$
Expansion (2020) [61]	Label Expansion through Finding Cores with Maximization of Similarity of Common Neighbors	Y	N	N	O	$O(n \log n)$
UMST (2020) [62]	Maximum Spanning Tree	Y	N	Y	O	$O(m \log k_{max})$
Ego Splitting (2017) [63]	Spanning Subgraphs of Node Neighbors	Y	N	Y	O	$O(m^{3/2})$
ANGEL (2020) [46]	Bottom-Up Clustering of Individual Spanning Subgraphs of Node Neighbors	Y	N	Y	O	$O(n)$
BIGCLAM (2013) [64]	Affiliation Strength Bipartite Network of Nodes and Communities using a Nonnegative Latent Factor Matrix	Y	N	N	O	$O(n)$
NNSED (2017) [65]	Encoding and Decoding Nonnegative Latent Factor Matrix	Y	N	Y	O	$O(kn^2)$

LE = Leading Eigenvector; CPM = Constant Potts Model; LP = Label Propagation; U = Undirected; D = Directed; W = Weighted; D/O = Disjoint/Overlapping; n = number of nodes; m = number of edges; i = number of iterations; k = node degree; c = number of communities; d = average node degree; q = size of cliques.

algorithms, it is imperative to establish a clear definition of a "community" within the context of the social graphs that we are constructing for the forums. A *community* refers to a subgraph C_i of graph G whose nodes are densely connected within C_i, but sparsely connected with nodes from other subgraphs C_j, where $j \neq i$ and $C_i \cap C_j = \emptyset$ for all i, j. Communities can be considered as dense clusters of nodes within the graph where frequently interacting users are physically closer to each other. Some communities could be considered overlapping, meaning that some nodes can claim membership to several partitions since they are involved in several densely interacting groups [33], in this case we drop the assumption that $C_i \cap C_j = \emptyset$.

We evaluated the performance of some recently developed community detection algorithms and those that have obtained benchmark status or have been validated on real-world social networks, as elaborated in the literature review (Section 2.2). Our assessment encompasses two distinct scenarios: the detection of disjoint communities and the identification of overlapping ones. The comprehensive list of algorithms used in this study and detailed information on the underlying assumptions that guide their methodologies is provided in Table 4. We utilize the direction of the link and the weights if the algorithm allows it; otherwise, the network is reduced to an undirected and unweighted graph. These algorithms are systematically categorized based on their primary approach and the underlying assumptions derived from the literature or our own bottom-up analysis.

We identified seven primary groups of algorithms. Random walk-based methods included InfoMap and Walktrap. Clustering-based methods comprised CPM, Surprise, and SCD. Modularity-based consisted of RB Potts, Louvain, Leiden, Leading Eigenvector, and EdMot. Propagation through neighbors encompassed Label Propagation, SLPA, and Expansion. Spanning trees-based included UMST. The Ego nets comprised Ego Splitting and ANGEL. Finally, nonnegative Latent Factor Matrix-based has two methods: BigClam and NNSED.

We expect algorithms within the same group to exhibit similar performance and substantial agreement on community labels. Furthermore, we also expect that modularity-based algorithms will achieve higher modularity scores. Algorithms that propagate through neighbors and those based on ego nets are expected to score higher in triplet hub potential. Overlapping community detection methods are likely to yield higher coverage scores because of the increased probability of nodes being assigned to the most fitting community, where they have fewer connections with the external network. The definition of the metrics is discussed in the next Section 3.4.

It is important to note that we have intentionally excluded algorithms that require a predefined number of clusters to search for, such as GEMSEC [66] or MNMF [67], as well as those that necessitate the specification of an initial set of nodes to initiate the search process, like Lswl-plus [68]. This exclusion is justified by the fact that, in the context of the problems addressed in this study, we lack prior knowledge regarding the approximate number of groups to identify or specific nodes to seed the grouping process. Furthermore, it is essential to recognize that in the domain of CTI, obtaining ground truth data is often unattainable and inherently challenging to derive, making the problem primarily unsupervised in nature.

3.4. Evaluation and analysis

To assess the performance of our algorithms, we employ metrics that closely align with the characteristics we aim to identify within the communities under investigation. These characteristics involve the presence of stronger connections within the communities compared to connections outside them, as well as a higher than expected distribution of connections within each community. Using a combination of measures can provide a better understanding of the algorithms' performance and quality of the detected communities. To evaluate the algorithms, we utilize three measures that are closely aligned with the characteristics we aim to uncover when investigating communities within CTI tasks. In addition to these standard measures, we introduce our quality metric, *triplet hub potential*, which we believe can help identify communities with potential dominating members. Using a combination of these measures, we can obtain a more comprehensive assessment of algorithmic performance.

Modularity measures the strength of a graph division into different clusters. A high modularity score is achieved when a network has densely populated communities with few links between them. A score near 1 indicates that community interaction is more separate and distinct, while a score near 0 indicates near-random member interaction and little community formation. To calculate modularity for directed weighted graphs, we use the function proposed by Nicosia et al. [69], and in the case of overlapping partitioning, we use the function by Lazar et al. [70].

However, a high modularity score does not always indicate the presence of a community structure in a network [49]. In random graphs, where there is no theoretical bias towards a particular clustering of nodes, high modularity clustering is still possible [11]. To address this issue, we also include another quality measure – *coverage*, which is defined as the fraction of intracommunity edges to the total number of edges in a graph [11]. In a perfect community structure where no two clusters are connected, coverage equals 1, since every edge in the graph is contained within a single community. We adapt the calculation of coverage for overlapping communities.

In addition to modularity and coverage, we use *performance* measure, which is defined as the ratio of the number of intracommunity edges plus intercommunity edges to the total number of *potential edges* (expected) in the graph [11]. To evaluate overlapping communities, in place of *performance*, we utilize a scoring metric known as the *normalized cut*. This metric, originally derived for image segmentation [71], but also used in the community detection problem [72], encapsulates the same logic as *performance* measure. It quantifies community quality by calculating the ratio of the sum of intercommunity edge weights to the sum of intracommunity edge weights for each identified community [71].

The fourth metric (our proposed method) is rooted in the concept of scale-free networks and builds upon the assumption that social networks exhibit a propensity for preferential attachment and the presence of existing hubs. To calculate the *triplet hub potential* measure, we focus on communities with three or more members. Within these communities, we identify the node with the highest degree (which encompasses both in-degree and out-degree) and then divide this degree by the potential maximum degree that the node could attain. A score of 1 denotes a perfect hub within the community, while a score approaching 0 indicates a community with no discernible hubs. We then calculated the average score for all communities identified by the algorithm. Below is the formal equation for computation.

The *Triplet Hub Potential* for a set of communities within a graph is defined as:

$$\text{Triplet Hub Potential} = \frac{1}{C} \sum_{i=1}^{C} \text{Triplet Hub Potential}_i, \tag{1}$$

where C represents the number of communities. The potential score Hub Potential_i for each community C_i is calculated as follows:

$$\text{Triplet Hub Potential}_i = \begin{cases} \frac{\max_{v \in C_i} (k_v^{out} + k_v^{in})}{2 \times (|C_i| - 1)}, & \text{if } |C_i| > 2 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where k_v^{out} and k_v^{in} are the out-degree and in-degree of a node v within the community C_i, respectively. The maximum is taken over all nodes v in the community C_i, and $|C_i|$ denotes the number of nodes in the community C_i. The Triplet Hub Potential score is computed only for communities that have more than two members. For communities with two or fewer members, the score is set to zero, which helps to penalize singletons and couples detected as communities.

The proposed measure is not necessarily intended to outperform other metrics. Rather, its purpose is to provide quantified insights into the nature of communities detected by algorithms. For example, while *modularity* assesses the potential for divisions within the network, the Triplet Hub (T-hub) Potential specifically quantifies the presence of highly connected nodes, or hubs, within these communities. In contrast to *coverage* and *performance*, which evaluate the distribution of connections between nodes and between communities, T-hub Potential focuses on quantifying the source of the most connections.

The *T-hub potential* metric also addresses the issue of over-segmentation by penalizing it. Grounded in the theory of preferential attachment, it recognizes communities that coalesce around key contributors. This approach is helpful in differentiating between egalitarian communities and those led by clear leadership or dominated by influential contributors. In Appendix, we show how the values of the score change on the theoretical graphs, such as star graph, complete graph, and others.

Finally, to comprehensively cross-compare the detected communities across different algorithms, we employ *omega index*. This statistical measure provides a pairwise agreement score between algorithms on their community labeling, including for overlapping partitions [73].

4. Results

Analysis of algorithms used to detect disjoint and overlapping communities in forums reveals varied performances. Some methods yield consistently similar results across different forum structures, while others show divergent behaviors depending on the specific forum. Fig. 2 presents these data through a series of radar (or spider) charts. In these graphs, each variable is represented on a separate axis radiating from a central point. The values are plotted along each axis and connected to form a polygon. Each chart corresponds to a selected forum (*antionline*, *exetools*, and *go4expert*), with separate analyzes for disjoint and overlap community detections. The density values are indicated in the title of the chart.

In these radar charts, algorithms that form a larger enclosed area indicate better performance in four measured metrics. For example, in the *exetools* forum with disjoint communities, the Leiden algorithm shows robust performance in all metrics, closely followed by the Louvain and RB Pots algorithms. Conversely, in the *go4expert* forum under similar conditions, algorithm performances are more uniform, likely due to the forum's sparsity and the presence of more disconnected components. Here, all algorithms demonstrate confidence in partitioning, and Leiden and Louvain excel in modularity. In particular, Louvain and Leiden consistently perform strongly in all five forums using four metrics. Clustering-based methods such as CPM, Surprise, and SCD score lower T-hub potential than others, indicating less effectiveness in identifying communities with node dominance.

The best performances across these metrics are summarized in Table 5, which highlights the five most effective algorithms. It should be noted that the algorithms, such as CPM, SCD, Surprise, UMST and Expansion, excelling in the *Normalized Cut* and *Performance* metrics also tend to identify a higher number of communities, as detailed in Table 6.

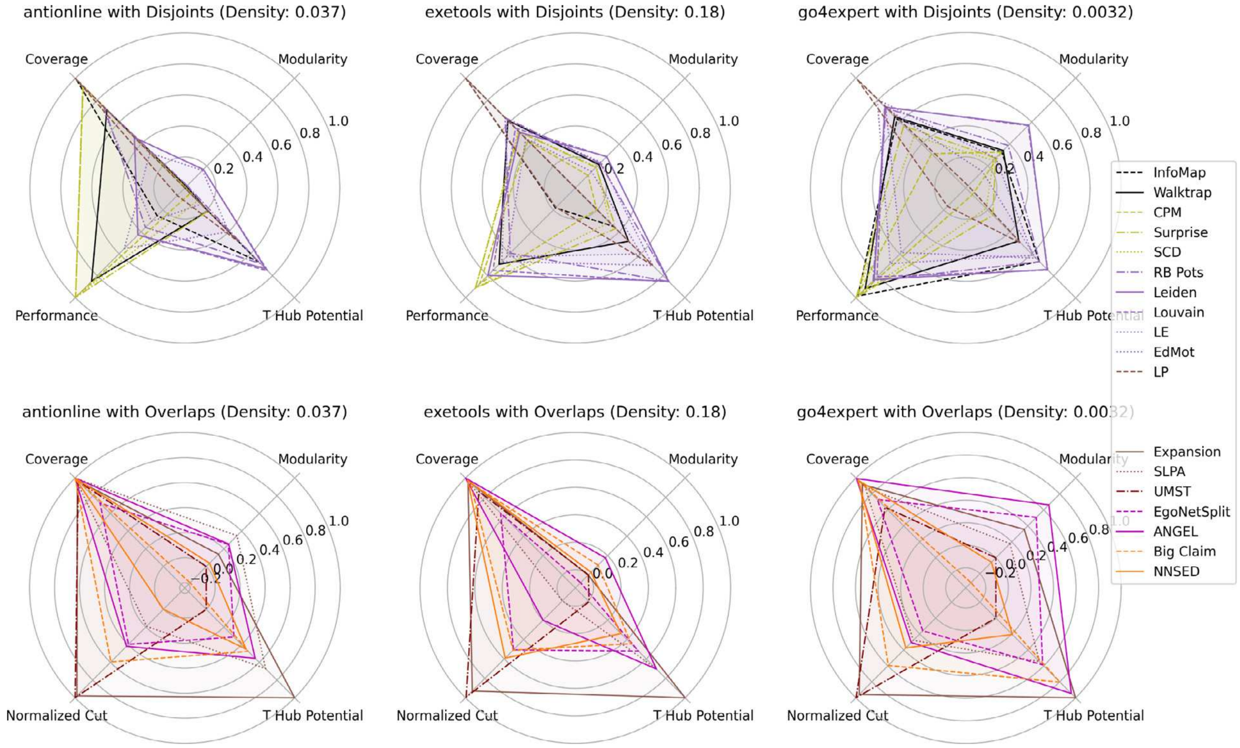


Fig. 2. *Modularity*, *Coverage*, *Performance* (or *Normalized Cut Ratio* for overlapping communities) and *Triplet Hub Potential* of disjoint and overlapping communities detections for selected forums (with densities).

For overlapping community detection, the Expansion algorithm, along with ANGEL and EgoNetSplitter algorithms based on ego-net, demonstrate superior performance in various forums. They particularly excel in terms of modularity and T-hub potential, as was expected due to the approach being based on ego nets. However, the UMST algorithm shows a consistent shortfall in maximizing these metrics. It does, however, stand out in terms of coverage and normalized cut metrics. The strong performance of the Expansion algorithm, especially in achieving high T-hub potential, is not unexpected. This is because the method inherently focuses on initially identifying hubs by searching for cores that maximize the similarity of common neighbors. Subsequently, it expands the subgraph using label propagation. This approach effectively uncovers influential nodes within communities, explaining its high T-hub potential scores. Generally, overlapping community detection methods are observed to yield superior quality metrics, likely due to their ability to flexibly assign nodes to multiple contextually relevant groups.

Table 6 presents the number of communities detected by various algorithms in five different forums. The data is arranged in descending order, revealing a pattern where algorithms with similar methodologies tend to be grouped together, suggesting that they likely identify similar community structures. For example, Louvain and Leiden algorithms, both designed for modularity optimization, yield comparable community counts across the forums. The clustering-based algorithms (CPM, SCD, Surprise) stand out for detecting a higher number of communities, suggesting that they are adept at uncovering more detailed community structures within the forums.

Conversely, algorithms that employ modularity optimization and neighborhood propagation strategies, including EgoNetSplitter, Louvain, and similar methods, tend to identify fewer communities across the forums, which could imply that these methods are possibly merging smaller communities into larger ones or identifying only the most prominent groups.

It should also be noted that algorithms like NNSED and BigClam consistently detect a lower number of communities across all forums, treating connected components (SCCs) as giant communities, which is not helpful for analysis. Note that these methods also do not score high in quality measures.

Fig. 3 shows the Omega index matrices for five distinct forums. Each matrix illustrates the pairwise correlation of community membership assignments by various community detection algorithms. These algorithms are grouped by approach and indicated through different font colors. The intensity of the color in each cell of the matrix reflects the Omega coefficient between the community memberships identified by the corresponding algorithms. The Omega coefficient ranges from −1 to 1, with 1 signifying perfect agreement, 0 indicating no correlation, and −1 representing perfect disagreement. Notably, in the *cardingteam* forum, algorithms tend to show more agreement in their findings (Average Omega Index = 0.209) compared to the other four forums. This could be attributed to the smaller size and relative sparsity of the *cardingteam* forum, which likely results in more clearly defined subgroups, as can be seen from the high scores for most algorithms. On the contrary, the *go4expert* forum shows a lower overall agreement among algorithms (Average Omega Index = 0.083), possibly due to its lower density and larger size, posing a challenge to the common methodologies of these algorithms.

In terms of clustering-based algorithms such as CPM, Surprise and SCD, a moderate to high level of agreement is observed, especially in the *exetools* and *cipher* forums. Algorithms such as RB Potts, Louvain, Leiden, LE, and EdMot, which focus on optimizing modularity, show strong internal agreement across all forums, with particularly high concordance in the *cardingteam*, *cipher*, and *antionline* forums. The Louvain and Leiden methods consistently exhibit close agreement. The group comprising algorithms that propagate through neighbors (Label Propagation, SLPA, Expansion) and ego-net-based methods (EgoNetSplitter, ANGEL) demonstrate varied levels of agreement. However, SLPA and LP, given their close relationship, show a relatively higher correlation with each other in multiple forums. Interestingly, the Expansion algorithm shows less agreement with the label propagation group but sometimes aligns closely with clustering-based algorithms. The UMST

Table 5
Top Five community detection algorithms ranked by median score across multiple forums. The font color coding represents the algorithmic groupings.

Algorithm	antionline	cardingteam	cipher	exetools	go4expert	Median
Top 5 by triplet hub potential						
Expansion	1.00000	1.00000	1.00000	1.00000	0.99932	1.00000
RB Pots	0.68836	0.82919	0.87185	0.85427	0.67199	0.82919
ANGEL	0.55386	0.77017	0.79067	0.70362	0.94548	0.77017
Louvain	0.75567	0.75150	0.80420	0.85649	0.74638	0.75567
Leiden	0.73549	0.68032	0.81085	0.84851	0.74204	0.74204
Top 5 by modularity						
Leiden	0.16905	0.57035	0.39124	0.28657	0.57392	0.39124
ANGEL	0.25392	0.56952	0.39001	0.17714	0.66376	0.39001
Louvain	0.16904	0.55605	0.38762	0.28511	0.56758	0.38762
EgoNetSplitter	0.25043	0.31118	0.32013	−0.07917	0.50680	0.31118
RB Pots	0.02098	0.61597	0.30558	0.22593	0.38652	0.30558
Top 5 by Performance/Normalized cut						
UMSTMO	0.99868	0.99875	0.99759	0.99335	0.99746	0.99759
SCD	0.99747	0.99076	0.98622	0.91044	0.99823	0.99076
Surprise	0.99735	0.97476	0.98604	0.91211	0.99690	0.98604
CPM	0.99727	0.98094	0.97907	0.87387	0.99769	0.98094
Expansion	0.97897	0.72909	0.97331	0.93012	0.95848	0.95848
Top 5 by coverage						
Label Propagation	0.99988	0.79205	0.99977	1.00000	0.98895	0.99977
ANGEL	0.99964	0.98013	0.99972	0.99997	0.99348	0.99964
NNSED	1.00000	0.68848	0.99371	0.97848	0.97460	0.97848
Expansion	0.97434	0.91038	0.98176	0.97447	0.91908	0.97434
SLPA	0.99996	0.81192	0.95263	1.00000	0.88478	0.95263

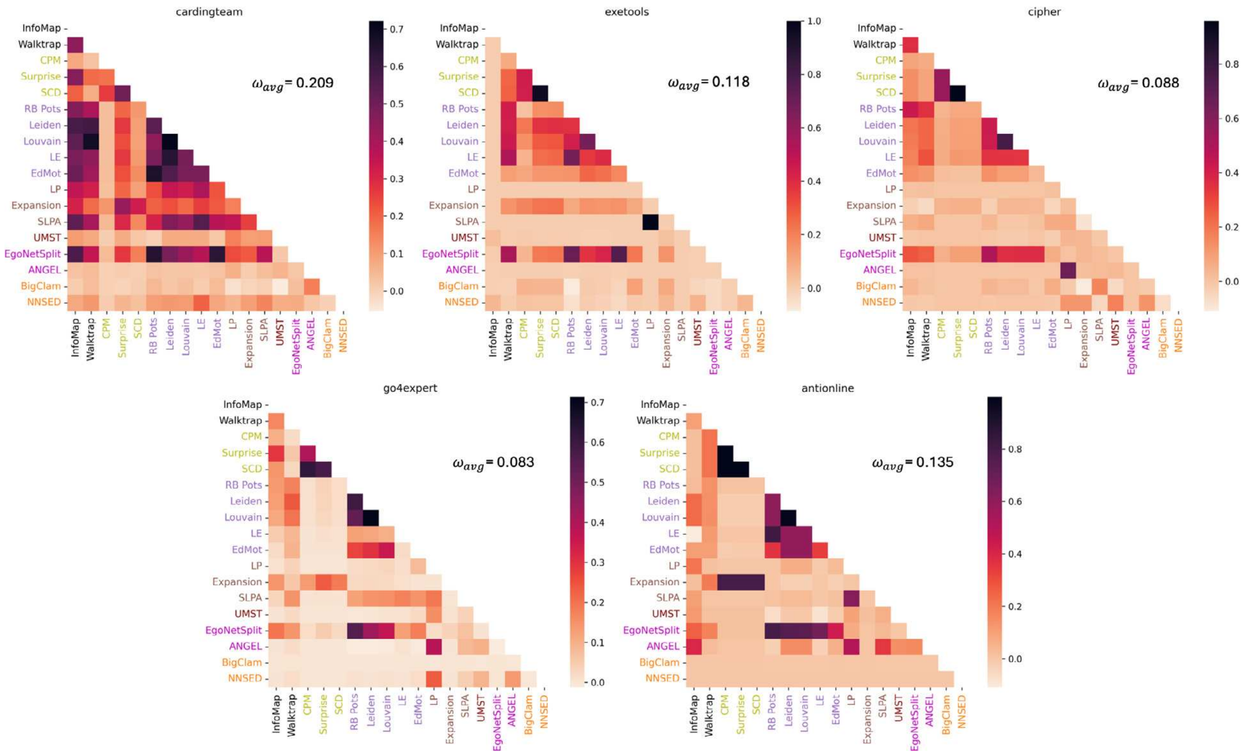


Fig. 3. Cross-Comparison of Community Detection Algorithms Across Forums Visualized with the *Omega Index*. The font color coding represents the algorithmic groupings. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

algorithm, based on spanning tree methods, tends to correlate less with most other algorithms. This divergence suggests a distinct approach to community detection and, as indicated in Fig. 2, it appears less effective overall in modularity and hub potential. EgoNetSplitter, surprisingly, agrees considerably with several non-overlapping community detection methods, particularly those based on modularity. This is intriguing given that EgoNetSplitter primarily uses local ego network structures and immediate neighbor labeling, while modularity-based algorithms optimize a global metric.

Fig. 4 visually illustrates the differences in community detection by three algorithms: Leiden, Expansion, and EgoNetSplitter. While all three of these algorithms achieve high scores in terms of modularity, EgoNetSplitter and Expansion stand out with their higher T-hub potential and coverage scores. This figure effectively highlights the

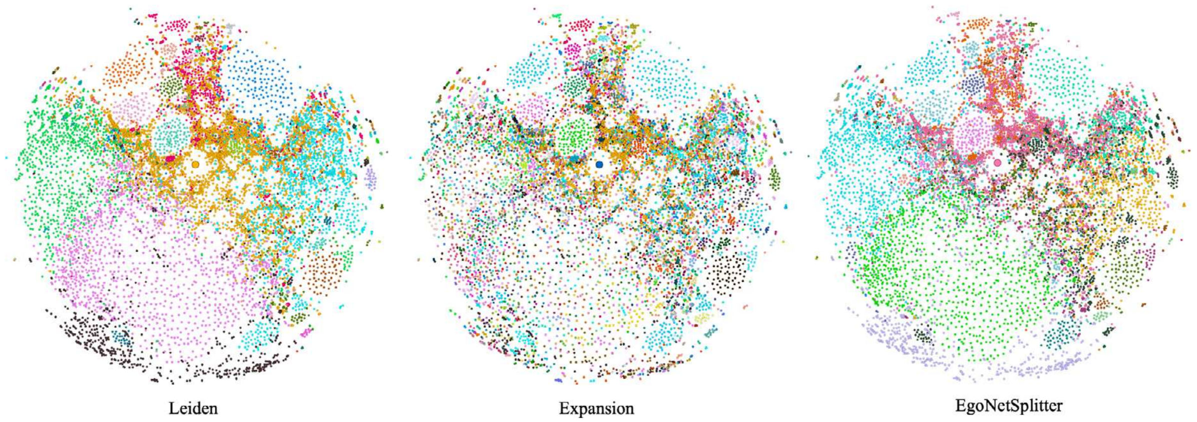


Fig. 4. Comparison of Communities detected in *go4expert* (edges are hidden).

Table 6
Number of communities detected by each algorithm. The font color coding represents the algorithmic groupings.

	antionline	cardingteam	cipher	exetools	go4expert
CPM	7439	342	1340	244	5849
SCD	5777	168	817	132	3747
Surprise	4651	123	750	146	2545
UMST	1235	211	258	80	2474
Expansion	771	60	183	59	986
Walktrap	801	17	51	5	505
Infomap	169	36	104	2	592
SLPA	26	31	12	1	233
Label Propagation	51	33	12	1	188
EdMot	54	15	13	5	46
EgoNetSplitter	28	13	18	5	41
Louvain	15	15	12	4	57
Leiden	14	14	12	5	53
Eigenvector	36	10	13	5	23
RB Pots	13	10	13	3	40
ANGEL	34	3	2	1	5
NNSED	1	4	2	3	9
BigClam	2	3	2	2	2

variances in community attribution among these algorithms, despite their shared success in modularity optimization. Although the Leiden and EgoNetSplitter algorithms identified a relatively similar number of communities (53 and 41, respectively), the Expansion algorithm detected a significantly higher number, with 986 smaller communities (the average size is 13), more intimate communities. This contrasts sharply with the average community sizes found by Leiden and EgoNetSplitter, which were 214 and 165, respectively. This difference highlights the unique approach of the Expansion algorithm in identifying smaller, more closely knit communities within the *go4expert* forum.

5. Discussion and conclusion

Exploring hacker communities where actors, including threat actors, form collaboration and develop emerging techniques is a research area that has garnered relatively little attention within academia. In this study, we investigate the realm of identifying optimal algorithms for detecting communal structures that can efficiently separate hackers with shared but implicit connections into distinct groups. To achieve this, we employ a comprehensive set of four quality metrics: *modularity*, *coverage*, *performance*, and the *normalized cut ratio*. These metrics serve as our guide as we navigate the intricacies of inferring the ground truth of these communities. In addition to these well-established metrics for community detection, we suggest a measure of community quality, the *Triplet hub potential*. This measure offers valuable information on the extent to which detected communities harbor the potential of influential leaders or leading actors. By incorporating this metric into our analysis, we gain a sense of the hierarchical dynamics within the identified communities.

Our investigation involved a thorough cross-comparison of community detection methods, categorizing algorithms based on the principles they operationalize. Generally, algorithms tend to concur on community partitions when they originate from the same methodological approach, although there are exceptions. We observed that two modularity-based methods, *Louvain* and *Leiden*, consistently performed optimally well, often identifying the same communities. In contrast, cluster-based approaches showed less efficacy. While the *Louvain* and *Leiden* methods are robust, overlapping approaches generally provided superior results, particularly ego-net-based approaches and the *Expansion* algorithm. The *Expansion* algorithm, in particular, tends to find smaller subcommunities, with a methodology characterized by initial core identification followed by community expansion through label propagation. This granularity is helpful for investigators aiming to uncover specific characteristics of community organizations. The ability of these algorithms to highlight smaller, preferentially attached communities enables a more targeted investigation, emphasizing the significance of selecting the appropriate algorithm to match the investigative focus.

Example case. We provide an example case that shows the nature of the Expansion algorithm for identifying smaller communities (Fig. 5 for the *exetools* forum). The large groupings of nodes show communities identified by Leiden, but smaller groupings (uniformly colored) are those identified by Expansion. Leiden identified five communities on 886 users of exetools, whereas Expansion – 59. This illustrates the tendency of the Expansion algorithm to choose smaller, tightly-knit sub-communities within the broader network. However, the agreement between the partitions chosen by Leiden and Expansion differs significantly, as indicated by the diverse node colors within a single community identified by Leiden.

We qualitatively analyzed one of the communities identified by Expansion (circled in Fig. 5). Several evident hubs are present within this community. One of the shared discussion threads (over 5000 comments) is dedicated to maintaining the open-source x64 debugger, a binary debugger for Windows developed for malware analysis and reverse engineering of executables without the source code. This open-source debugger is verified by the security community and sponsored by several entities. However, due to the nature of the software, it can also be used for malicious purposes, such as reverse engineering proprietary software or analyzing the behavior of a produced malicious software.

One of the most connected members is the developer of the tool, while at least three others actively provide feedback, suggest improvements, and offer encouragement. Another member is also associated with the tool. All of these users were identified as belonging to the same community by Expansion and different communities – by Leiden. While we do not share the usernames of the members, here is a partial quote from the major developer identified in the community:

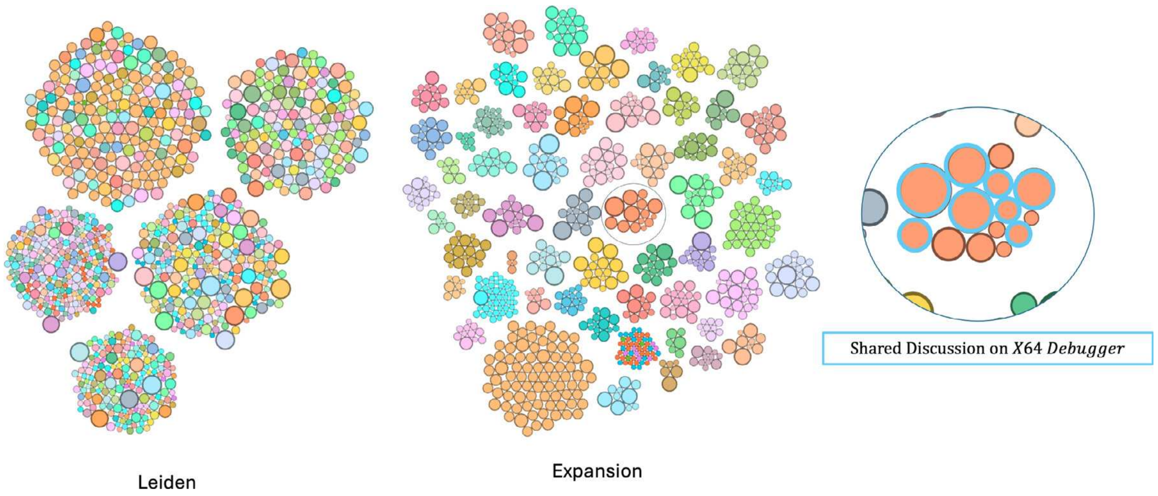


Fig. 5. Comparison of communities detected in *exetools* (edges are hidden). Larger clusters are identified by *Leiden*, while smaller, uniformly colored clusters are identified by *Expansion*. The close-up shows the selected community for qualitative analysis. The highlighted nodes (blue) contributed to discussions about the x64 debugger. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

> "Please provide as much information as you can on crashes. Please also try the snapshots found here to see if certain issues are maybe already fixed..."

Here is a selected response from another member of the detected community:

> "What can i say other than very nice work, finally someone to pick up the thread on x64. What I do wonder is though if you can implement a feature so that we can be able to search full memory..."

Other shared discussion threads had themes related to a disassembler, a debugger, and other tools that show binary instructions that a processor executes.

Suggested approach. Drawing on the findings of this study, we propose an approach for CTI and OSINT researchers to employ community detection algorithms. By understanding the shared theoretical foundations and recognizing the distinctions in the outcomes produced by these algorithms across various metrics, researchers can better interpret the nature of hacker communities. Here, we suggest a guide for CTI and OSINT researchers to conduct investigative analysis using community detection benchmark methods. The analysis of the collected dataset from a hacker forum can be exploratory but will aid researchers in a more focused exploration of the user content and interactions on the forums. Researchers should consider the following:

- Use multiple community detection methods from different methodological approaches. As we have seen, models rarely agree on partitioning and mostly agree if the core methodological approach is the same.
- Use benchmarks such as *Leiden* or *Louvain*. Both of these models are validated in numerous experiments, including this study, and yield high results in quality measures.
- Use multiple metrics to assess the quality of detected communities, including a proposed metric, *Triplet Hub Potential*. This metric indicates the presence of a hub in the community, ignoring node couples found as a community, which, along with *Modularity*, *Coverage*, and *Performance*, provides a comprehensive overview of the nature and structure of the detected communities.
- Use the *Expansion* algorithm or other methods based on propagation through the neighbors of nodes. Our results suggest that the *Expansion* algorithm excels in performance and also finds smaller, nuanced communities with notable hubs. Such scoped-sized communities are more manageable for researchers to investigate.
- Investigate identified communities through shared contributions to specific thread discussions. We provided an example case that delves into the analysis of a single community identified by Expansion.

In summary, methods that investigate the neighborhood characteristics of nodes tend to yield groupings that are smaller and potentially easier to investigate. This highlights the potential of core expansions and neighbor-propagation-based approaches as a fruitful avenue for the development of community detection techniques tailored to cyber threat intelligence tasks, especially if there is a specific user that investigators are monitoring already. Furthermore, this study has led to the creation of an interactive tool that facilitates the exploration of community structures. It allows users to apply various detection algorithms and employ visual cues to navigate detected clusters.[1] Ultimately, such insights and approaches can play an important role for researchers seeking to develop advanced or targeted cyber threat intelligence research capabilities.

CRediT authorship contribution statement

Dalyapraz Manatova: Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Charles DeVries:** Software. **Sagar Samtani:** Supervision, Data curation, Conceptualization.

Declaration of competing interest

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

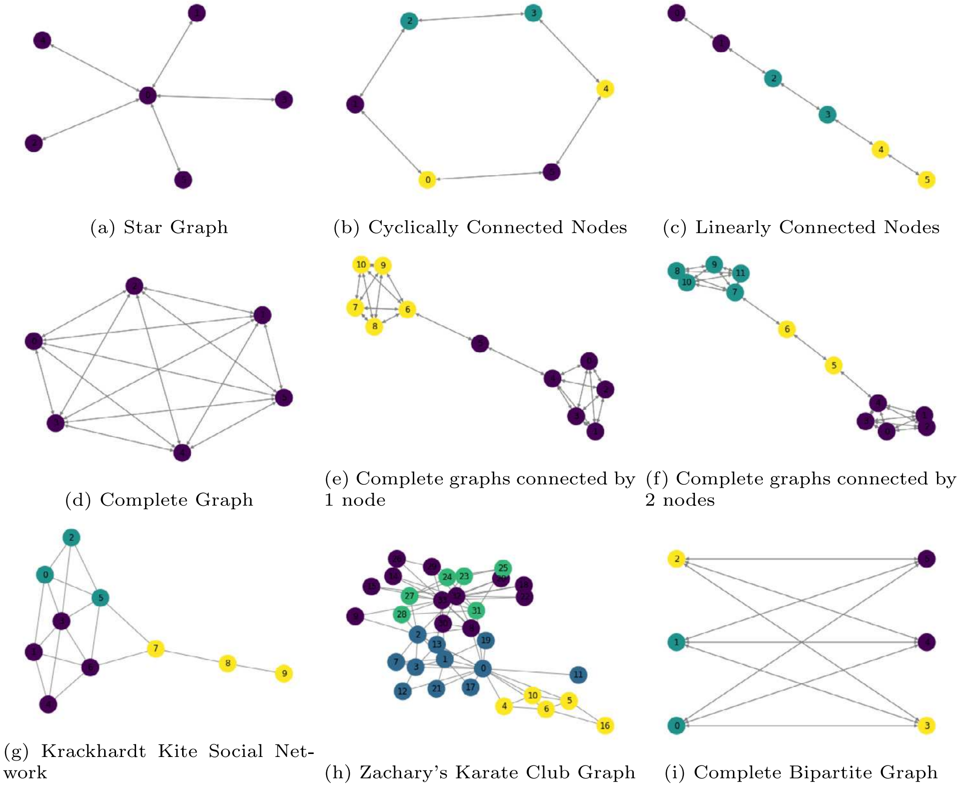
Data availability

Data will be made available on request.

Acknowledgments

This material is based upon work supported by the National Science Foundation, USA under Grant No. OAC-1917117, CNS-2338479, DGE-1946537, OAC-2319325, and CNS-1850362.

[1] https://go.iu.edu/hackerforum-community-detection



(a) Star Graph (b) Cyclically Connected Nodes (c) Linearly Connected Nodes

(d) Complete Graph (e) Complete graphs connected by 1 node (f) Complete graphs connected by 2 nodes

(g) Krackhardt Kite Social Network (h) Zachary's Karate Club Graph (i) Complete Bipartite Graph

Fig. A.6. Communities Detected on Theoretical Graphs using Leiden.

Table A.7
Results of community detection metrics for Fig. A.6.

Figure	Modularity	Coverage	Performance	T-hub mean	T-hub values
a	0.0	1.0	0.333	1.0	[1.0]
b	−0.167	0.167	0.533	0.0	[0, 0, 0]
c	0.26	0.6	0.867	0.0	[0, 0, 0]
d	0.0	1.0	1.0	1.0	[1.0]
e	0.454	0.955	0.909	1.0	[1.0, 1.0]
f	0.489	0.913	0.97	0.667	[1.0, 1.0, 0]
g	0.222	0.611	0.822	0.5	[0.5, 0.5, 0.5]
h	0.42	0.731	0.804	0.419	[0.5, 0.5, 0.3, 0.375]
i	−0.222	0.111	0.333	0.0	[0, 0, 0]

Appendix. Triplet hub potential results on theoretical graphs

Table A.7 presents the Triplet Hub Potential results for detected communities in the theoretical graphs depicted in Fig. A.6. Each theoretical graph exemplifies a specific extreme structure, such as star, cycle, bipartite complete, and fully complete graphs. Additionally, we include two well-established social graphs frequently used in network science: Krackhardt's organizational social graph [74] and Zachary's Karate Club social graph [75].

Communities within these graphs were detected using the Leiden algorithm, with each node's community membership indicated by color. We calculated four metrics: modularity, coverage, performance, and T-hub Potential. The values contributing to the average T-hub Potential are also provided.

The proposed metric assigns a score of 1 for cases of perfect hub existence, such as in Star and Complete Graphs, where all nodes function as perfect hubs. Conversely, in cases like Line, Cycle, and Bipartite graphs, where communities are divided into nodes with few interlinks or less than 3 nodes, the score is 0. Also note the change in partitioning and score as complete two graphs connected by one node (Fig. A.6(e)) change to the complete two graphs connected by two nodes (Fig. A.6(f)).

References

[1] J. Lewis, Economic Impact of Cybercrime, Tech. rep., The Center for Strategic and International Studies (CSIS), McAfee, 2018, URL http://csis-website-prod.s3.amazonaws.com/s3fs-public/publication/economic-impact-cybercrime.pdf.

[2] J. Lusthaus, How organised is organised cybercrime? Glob. Crime 14 (1) (2013) 52–60, Publisher: Routledge _eprint: https://doi.org/10.1080/17440572.2012.759508.

[3] M. Yip, N. Shadbolt, C. Webber, Structural analysis of online criminal social networks, in: 2012 IEEE International Conference on Intelligence and Security Informatics, 2012, pp. 60–65, http://dx.doi.org/10.1109/ISI.2012.6284092.

[4] D. Manatova, D. Sharma, S. Samtani, L.J. Camp, Building and testing a network of social trust in an underground forum: Robust connections and overlapping criminal domains, in: 2022 APWG Symposium on Electronic Crime Research, ECrime, 2022, pp. 1–12, http://dx.doi.org/10.1109/eCrime57793.2022.10142120, ISSN: 2159-1245.

[5] D. Manatova, L.J. Camp, J.R. Fox, S. Kuebler, M.A. Shardakova, I. Kouper, An Argument for Linguistic Expertise in Cyberthreat Analysis: LOLSec in Russian Language eCrime Landscape, IEEE Computer Society, 2023, pp. 170–176, http://dx.doi.org/10.1109/EuroSPW59978.2023.00024, URL https://www.computer.org/csdl/proceedings-article/eurospw/2023/272000a170/1OFtfig8SyI, ISSN: 2768-0657.

[6] V. Garg, L.J. Camp, Why cybercrime? ACM SIGCAS Comput. Soc. 45 (2) (2015) 20–28, http://dx.doi.org/10.1145/2809957.2809962, URL https://dl.acm.org/doi/10.1145/2809957.2809962.

[7] B. Collier, R. Clayton, A. Hutchings, D. Thomas, Cybercrime is (often) boring: Infrastructure and alienation in a deviant subculture, Brit. J. Criminol. 61 (5) (2021) 1407–1423, http://dx.doi.org/10.1093/bjc/azab026.

[8] J. Lusthaus, Honour Among (Cyber)thieves? Eur. J. Sociol. / Arch. Eur. Sociol. 59 (2) (2018) 191–223, http://dx.doi.org/10.1017/S0003975618000115, URL https://www.cambridge.org/core/journals/european-journal-of-sociology-archives-europeennes-de-sociologie/article/honour-among-cyberthieves/4B1CBA1B4F8AFC05FC7CD2BD8E44EFE7, Publisher: Cambridge University Press.

[9] R. Broadhurst, P. Grabosky, M. Alazab, S. Chon, Organizations and Cyber crime: An Analysis of the Nature of Groups engaged in Cyber Crime, Int. J. Cyber Criminol. 8 (1) (2014) 1–20, URL https://www.proquest.com/docview/1545341663/abstract/99B087197F9F4ACCPQ/1, Num Pages: 20 Place: Thirunelveli, India Publisher: International Journal of Cyber Criminology.

[10] E.R. Leukfeldt, E.R. Kleemans, W.P. Stol, Cybercriminal networks, social ties and online forums: Social ties versus digital ties within phishing and malware networks, Brit. J. Criminol. 57 (3) (2017) 704–722, http://dx.doi.org/10.1093/bjc/azw009.

[11] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3) (2010) 75–174, http://dx.doi.org/10.1016/j.physrep.2009.11.002, URL https://www.sciencedirect.com/science/article/pii/S0370157309002841.

[12] T. Pourhabibi, K.-L. Ong, B.H. Kam, Y.L. Boo, DarkNetExplorer (DNE): Exploring dark multi-layer networks beyond the resolution limit, Decis. Support Syst. 146 (2021) 113537, http://dx.doi.org/10.1016/j.dss.2021.113537, URL https://www.sciencedirect.com/science/article/pii/S0167923621000476.

[13] J. Tachaiya, J. Gharibshah, E.E. Papalexakis, M. Faloutsos, RThread: A thread-centric analysis of security forums, in: Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2020, 2020, pp. 473–477.

[14] I. Pete, J. Hughes, Y.T. Chua, M. Bada, A social network analysis and comparison of six dark web forums, in: Proceedings - 5th IEEE European Symposium on Security and Privacy Workshops, Euro S and PW 2020, 2020, pp. 484–493.

[15] S. Sarkar, M. Almukaynizi, J. Shakarian, P. Shakarian, Mining user interaction patterns in the darkweb to predict enterprise cyber incidents, Soc. Netw. Anal. Min. 9 (1) (2019).

[16] S.-Y. Huang, T. Ban, A topic-based unsupervised learning approach for online underground market exploration, in: 2019 18th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/13th IEEE International Conference on Big Data Science and Engineering, TrustCom/BigDataSE, 2019, pp. 208–215, http://dx.doi.org/10.1109/TrustCom/BigDataSE.2019.00036.

[17] E. Marin, M. Almukaynizi, E. Nunes, P. Shakarian, Community finding of malware and exploit vendors on darkweb marketplaces, in: 2018 1st International Conference on Data Intelligence and Security, ICDIS, 2018, pp. 81–84, http://dx.doi.org/10.1109/ICDIS.2018.00019.

[18] S.-Y. Huang, H. Chen, Exploring the online underground marketplaces through topic-based social network and clustering, in: 2016 IEEE Conference on Intelligence and Security Informatics, ISI, 2016, pp. 145–150, http://dx.doi.org/10.1109/ISI.2016.7745458.

[19] S. Samtani, M. Kantarcioglu, H. Chen, Trailblazing the artificial intelligence for cybersecurity discipline: A multi-disciplinary research roadmap, ACM Trans. Manage. Inf. Syst. 11 (4) (2020) http://dx.doi.org/10.1145/3430360, Place: New York, NY, USA Publisher: Association for Computing Machinery.

[20] P.Y. Du, N. Zhang, M. Ebrahimi, S. Samtani, B. Lazarine, N. Arnold, R. Dunn, S. Suntwal, G. Angeles, R. Schweitzer, H. Chen, Identifying, collecting, and presenting hacker community data: Forums, IRC, carding shops, and DNMs, in: 2018 IEEE International Conference on Intelligence and Security Informatics, ISI 2018, 2018, pp. 70–75.

[21] W. Li, H. Chen, J.F. Nunamaker, Identifying and profiling key sellers in cyber carding community: AZSecure text mining system, J. Manage. Inf. Syst. 33 (4) (2016) 1059–1086.

[22] J. Grisham, S. Samtani, M. Patton, H. Chen, Identifying mobile malware and key threat actors in online hacker forums for proactive cyber threat intelligence, in: 2017 IEEE International Conference on Intelligence and Security Informatics: Security and Big Data, ISI 2017, 2017, pp. 13–18.

[23] Y. Zhang, Y. Fan, Y. Ye, L. Zhao, J. Wang, Q. Xiong, F. Shao, KADetector: Automatic identification of key actors in online hack forums based on structured heterogeneous information network, in: 2018 IEEE International Conference on Big Knowledge, ICBK, 2018, pp. 154–161, http://dx.doi.org/10.1109/ICBK.2018.00028.

[24] E. Marin, J. Shakarian, P. Shakarian, Mining Key-Hackers on Darkweb Forums, in: 2018 1st International Conference on Data Intelligence and Security, ICDIS, 2018, pp. 73–80, http://dx.doi.org/10.1109/ICDIS.2018.00018.

[25] C. Huang, Y. Guo, W. Guo, Y. Li, HackerRank: Identifying key hackers in underground forums, Int. J. Distrib. Sens. Netw. 17 (5) (2021).

[26] J. Gharibshah, M. Faloutsos, Extracting actionable information from security forums, in: Companion Proceedings of the 2019 World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 27–32, http://dx.doi.org/10.1145/3308560.3314197, event-place: San Francisco, USA.

[27] N. Arnold, M. Ebrahimi, N. Zhang, B. Lazarine, M. Patton, H. Chen, S. Samtani, Dark-net ecosystem cyber-threat intelligence (CTI) tool, in: 2019 IEEE International Conference on Intelligence and Security Informatics, ISI, 2019, pp. 92–97, http://dx.doi.org/10.1109/ISI.2019.8823501.

[28] I. Deliu, C. Leichter, K. Franke, Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks, in: 2017 IEEE International Conference on Big Data, Big Data, 2017, pp. 3648–3656, http://dx.doi.org/10.1109/BigData.2017.8258359.

[29] Y. Zhang, Y. Fan, S. Hou, J. Liu, Y. Ye, T. Bourlai, IDetector: Automate underground forum analysis based on heterogeneous information network, in: Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '18, IEEE Press, 2018, pp. 1071–1078, event-place: Barcelona, Spain.

[30] R. Williams, S. Samtani, M. Patton, H. Chen, Incremental hacker forum exploit collection and classification for proactive cyber threat intelligence: An exploratory study, in: 2018 IEEE International Conference on Intelligence and Security Informatics, ISI, 2018, pp. 94–99, http://dx.doi.org/10.1109/ISI.2018.8587336.

[31] B. Biswas, A. Mukhopadhyay, S. Bhattacharjee, A. Kumar, D. Delen, A text-mining based cyber-risk assessment and mitigation framework for critical analysis of online hacker forums, Decis. Support Syst. 152 (2022) 113651, http://dx.doi.org/10.1016/j.dss.2021.113651, URL https://www.sciencedirect.com/science/article/pii/S0167923621001615.

[32] K. Otto, B. Ampel, S. Samtani, H. Zhu, H. Chen, Exploring the evolution of exploit-sharing hackers: An unsupervised graph embedding approach, in: 2021 IEEE International Conference on Intelligence and Security Informatics, ISI, 2021, pp. 1–6, http://dx.doi.org/10.1109/ISI53945.2021.9624846.

[33] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Defining and identifying communities in networks, Proc. Natl. Acad. Sci. 101 (9) (2004) 2658–2663, http://dx.doi.org/10.1073/pnas.0400054101, URL https://www.pnas.org/doi/10.1073/pnas.0400054101, Publisher: Proceedings of the National Academy of Sciences.

[34] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. 99 (12) (2002) 7821–7826, http://dx.doi.org/10.1073/pnas.122653799, URL https://www.pnas.org/doi/full/10.1073/pnas.122653799, Publisher: Proceedings of the National Academy of Sciences.

[35] P. Sah, L.O. Singh, A. Clauset, S. Bansal, Exploring community structure in biological networks with random graphs, BMC Bioinform. 15 (1) (2014) 220, http://dx.doi.org/10.1186/1471-2105-15-220.

[36] M.A. Javed, M.S. Younis, S. Latif, J. Qadir, A. Baig, Community detection in networks: A multidisciplinary review, J. Netw. Comput. Appl. 108 (2018) 87–111, http://dx.doi.org/10.1016/j.jnca.2018.02.011, URL https://www.sciencedirect.com/science/article/pii/S1084804518300560.

[37] O. Elezaj, S.Y. Yayilgan, E. Kalemi, Criminal network community detection in social media forensics, in: S. Yildirim Yayilgan, I.S. Bajwa, F. Sanfilippo (Eds.), Intelligent Technologies and Applications, in: Communications in Computer and Information Science, Springer International Publishing, Cham, 2021, pp. 371–383, http://dx.doi.org/10.1007/978-3-030-71711-7_31.

[38] A. Lancichinetti, F. Radicchi, J.J. Ramasco, S. Fortunato, Finding statistically significant communities in networks, PLoS One 6 (4) (2011) e18961, http://dx.doi.org/10.1371/journal.pone.0018961, URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0018961, Publisher: Public Library of Science.

[39] A. Prat-Pérez, D. Dominguez-Sal, J.-L. Larriba-Pey, High quality, scalable and parallel community detection for large real graphs, in: Proceedings of the 23rd International Conference on World Wide Web, WWW '14, Association for Computing Machinery, New York, NY, USA, 2014, pp. 225–236, event-place: Seoul, Korea.

[40] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. Theory Exp. 2008 (10) (2008) P10008.

[41] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 026113, http://dx.doi.org/10.1103/PhysRevE.69.026113, URL https://link.aps.org/doi/10.1103/PhysRevE.69.026113, Publisher: American Physical Society.

[42] V.A. Traag, L. Waltman, N.J. van Eck, From Louvain to Leiden: guaranteeing well-connected communities, Sci. Rep. 9 (1) (2019) Publisher: Nature Publishing Group.

[43] A. Clauset, M.E.J. Newman, C. Moore, Finding community structure in very large networks, Phys. Rev. E Stat. Nonlinear Soft Matter Phys. 70 (6 Pt 2) (2004) 066111.

[44] P. Pons, M. Latapy, Computing communities in large networks using random walks, in: Computer and Information Sciences - ISCIS 2005, Springer Berlin Heidelberg, 2005, pp. 284–293.

[45] M. Rosvall, C.T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proc. Natl. Acad. Sci. USA 105 (4) (2008) 1118–1123, Publisher: National Academy of Sciences.

[46] G. Rossetti, ANGEL: efficient, and effective, node-centric community discovery in static and dynamic networks, Appl. Netw. Sci. 5 (1) (2020) 1–23, http://dx.doi.org/10.1007/s41109-020-00270-6, URL https://appliednetsci.springeropen.com/articles/10.1007/s41109-020-00270-6, Number: 1 Publisher: SpringerOpen.

[47] Z. Zhang, Q. Li, D. Zeng, H. Gao, User community discovery from multi-relational networks, Decis. Support Syst. 54 (2) (2013) 870–879, http://dx.doi.org/10.1016/j.dss.2012.09.012, URL https://www.sciencedirect.com/science/article/pii/S0167923612002473.

[48] F. Liu, S. Xue, J. Wu, C. Zhou, W. Hu, C. Paris, S. Nepal, J. Yang, P.S. Yu, Deep learning for community detection: progress, challenges and opportunities, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI '20, Yokohama, Yokohama, Japan, 2021, pp. 4981–4987.

[49] R. Guimerà, M. Sales-Pardo, L.A.N. Amaral, Modularity from fluctuations in random graphs and complex networks, Phys. Rev. E 70 (2) (2004) 025101, http://dx.doi.org/10.1103/PhysRevE.70.025101, URL https://link.aps.org/doi/10.1103/PhysRevE.70.025101, Publisher: American Physical Society.

[50] J. Raber, Columbo: High perfomance unpacking, in: 2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering, SANER, 2017, pp. 507–510, http://dx.doi.org/10.1109/SANER.2017.7884663.

[51] M. Al-Ramahi, I. Alsmadi, J. Davenport, Exploring hackers assets: Topics of interest as indicators of compromise, in: Proceedings of the 7th Symposium on Hot Topics in the Science of Security, HotSoS '20, Association for Computing Machinery, New York, NY, USA, 2020, http://dx.doi.org/10.1145/3384217.3385619, event-place: Lawrence, Kansas.

[52] S. Pastrana, A. Hutchings, A. Caines, P. Buttery, Characterizing eve: Analysing cybercrime actors in a large underground forum, in: M. Bailey, T. Holz, M. Stamatogiannakis, S. Ioannidis (Eds.), Research in Attacks, Intrusions, and Defenses, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2018, pp. 207–227, http://dx.doi.org/10.1007/978-3-030-00470-5_10.

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

[54] V.A. Traag, P. Van Dooren, Y. Nesterov, Narrow scope for resolution-limit-free community detection, Phys. Rev. E 84 (1) (2011) 016114, http://dx.doi.org/10.1103/PhysRevE.84.016114, URL https://link.aps.org/doi/10.1103/PhysRevE.84.016114, Publisher: American Physical Society.

[55] V.A. Traag, R. Aldecoa, J.-C. Delvenne, Detecting communities using asymptotical surprise, Phys. Rev. E 92 (2) (2015) 022816, http://dx.doi.org/10.1103/PhysRevE.92.022816, URL https://link.aps.org/doi/10.1103/PhysRevE.92.022816, Publisher: American Physical Society.

[56] J. Reichardt, S. Bornholdt, Statistical mechanics of community detection, Phys. Rev. E 74 (1) (2006) 016110, http://dx.doi.org/10.1103/PhysRevE.74.016110, URL https://link.aps.org/doi/10.1103/PhysRevE.74.016110, Publisher: American Physical Society.

[57] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (3) (2006) 036104, Publisher: American Physical Society.

[58] P.-Z. Li, L. Huang, C.-D. Wang, J.-H. Lai, EdMot: An edge enhancement approach for motif-aware community detection, 2019, _eprint: 1906.04560.

[59] G. Cordasco, L. Gargano, Community detection via semi-synchronous label propagation algorithms, 2011, http://dx.doi.org/10.1504/..045103, URL https://arxiv.org/abs/1103.4550v1.

[60] J. Xie, B.K. Szymanski, X. Liu, SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process, in: 2011 IEEE 11th International Conference on Data Mining Workshops, 2011, pp. 344–349, http://dx.doi.org/10.1109/ICDMW.2011.154, URL https://ieeexplore.ieee.org/document/6137400, ISSN: 2375-9259.

[61] A. Choumane, A. Awada, A. Harkous, Core expansion: A new community detection algorithm based on neighborhood overlap, Soc. Netw. Anal. Min. 10 (1) (2020) 30, http://dx.doi.org/10.1007/s13278-020-00647-6.

[62] K. Asmi, D. Lotfi, M. El Marraki, Overlapping community detection based on the union of all maximum spanning trees, Libr. Hi Tech 38 (2) (2020) 276–292, http://dx.doi.org/10.1108/LHT-01-2019-0003, Publisher: Emerald Publishing Limited.

[63] A. Epasto, S. Lattanzi, R. Paes Leme, Ego-splitting framework: from non-overlapping to overlapping clusters, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 145–154, http://dx.doi.org/10.1145/3097983.3098054, URL https://dl.acm.org/doi/10.1145/3097983.3098054.

[64] J. Yang, J. Leskovec, Overlapping community detection at scale: a nonnegative matrix factorization approach, in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13, Association for Computing Machinery, New York, NY, USA, 2013, pp. 587–596, http://dx.doi.org/10.1145/2433396.2433471, URL https://dl.acm.org/doi/10.1145/2433396.2433471.

[65] B.-J. Sun, H. Shen, J. Gao, W. Ouyang, X. Cheng, A non-negative symmetric encoder-decoder approach for community detection, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 597–606, http://dx.doi.org/10.1145/3132847.3132902, URL https://dl.acm.org/doi/10.1145/3132847.3132902.

[66] B. Rozemberczki, R. Davies, R. Sarkar, C. Sutton, GEMSEC: Graph embedding with self clustering, 2019, _eprint: 1802.03997.

[67] X. Wang, P. Cui, J. Wang, J. Pei, W. Zhu, S. Yang, Community preserving network embedding, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI '17, AAAI Press, San Francisco, California, USA, 2017, pp. 203–209.

[68] S.E. Ayeb, B. Hemery, F. Jeanne, C. Charrier, E. Cherrier, Multigraph Transformation for Community Detection Applied to Financial Services, in: Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '22, IEEE Press, Istanbul, Turkey, 2023, pp. 541–548, http://dx.doi.org/10.1109/ASONAM55673.2022.10068607, URL https://dl.acm.org/doi/10.1109/ASONAM55673.2022.10068607.

[69] V. Nicosia, G. Mangioni, V. Carchiolo, M. Malgeri, Extending the definition of modularity to directed graphs with overlapping communities, J. Stat. Mech. Theory Exp. 2009 (03) (2009) P03024, http://dx.doi.org/10.1088/1742-5468/2009/03/P03024.

[70] A. Lázár, D. Ábel, T. Vicsek, Modularity measure of networks with overlapping communities, Europhys. Lett. 90 (1) (2010) 18001, http://dx.doi.org/10.1209/0295-5075/90/18001, URL http://arxiv.org/abs/0910.5072, arXiv:0910.5072 [physics].

[71] J.-b. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. (2000) URL https://www.academia.edu/29932002/Normalized_cuts_and_image_segmentation.

[72] W. Lin, M. Li, S. Zhou, J. Liu, G. Chen, Z. Gu, Phase transitions in normalized cut of social networks, Phys. Lett. A 383 (25) (2019) 3037–3042, http://dx.doi.org/10.1016/j.physleta.2019.06.042, URL https://www.sciencedirect.com/science/article/pii/S037596011930578X.

[73] J. Xie, S. Kelley, B.K. Szymanski, Overlapping community detection in networks: The state-of-the-art and comparative study, ACM Comput. Surv. 45 (4) (2013) 43:1–43:35, http://dx.doi.org/10.1145/2501654.2501657, URL https://dl.acm.org/doi/10.1145/2501654.2501657.

[74] D. Krackhardt, Assessing the political landscape: Structure, cognition, and power in organizations, Adm. Sci. Q. 35 (2) (1990) 342–369, http://dx.doi.org/10.2307/2393394, URL https://www.jstor.org/stable/2393394, Publisher: [Sage Publications, Inc., Johnson Graduate School of Management, Cornell University].

[75] W.W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol. Res. 33 (4) (1977) 452–473, http://dx.doi.org/10.1086/jar.33.4.3629752, URL https://www.journals.uchicago.edu/doi/10.1086/jar.33.4.3629752, Publisher: The University of Chicago Press.

Dalyapraz Manatova, a Ph.D. candidate in Security Informatics at Indiana University, focuses on the organizational and social dynamics of cybercriminals and outlawed communities. Dalya is also an Ostrom Fellow. With degrees from California State University, Los Angeles, and Nazarbayev University, Dalya has published and presented research on ecrime forums, cybercriminal groups, and vulnerability management tools at various conferences, including BlackHat, M3AAWG, IEEE European Symposium on Security and Privacy, APWG eCrime Symposium and PEARC.

Dr. Sagar Samtani is an Assistant Professor and Grant Thornton Scholar at Indiana University's Kelley School of Business. He earned his Ph.D. in 2018 from the University of Arizona's AI Lab. His research focuses on AI for cybersecurity, cyber threat intelligence, and Dark Web analytics, with expertise in deep learning, smart vulnerability assessment, and IoT. Dr. Samtani has held editorial positions at ACM Digital Threats and ACM Transactions on Management Information Systems. He has received numerous awards and grants from the National Science Foundation, including the prestigious IU Outstanding Junior Faculty Award in 2023. His publications span esteemed journals such as the Journal of Management Information Systems, MIS Quarterly, and IEEE Transactions on Dependable and Secure Computing.