Yeung et al. *EPJ Data Science* (2025) 14:37
https://doi.org/10.1140/epjds/s13688-025-00553-x

EPJ Data Science
a SpringerOpen Journal

RESEARCH Open Access

Garbage in garbage out? Impacts of data quality on criminal network intervention

Wang Ngai Yeung[1,3], Riccardo Di Clemente[1,4] and Renaud Lambiotte[2*]

Handling Editor: Daniel Romero

*Correspondence:
renaud.lambiotte@maths.ox.ac.uk
[2]Mathematical Institute, University of Oxford, Andrew Wiles Building, Oxford, OX2 6GG, UK
Full list of author information is available at the end of the article

Abstract

Criminal networks such as human trafficking rings are threats to the rule of law, democracy and public safety in our global society. Network science provides invaluable tools to identify key players and design interventions for Law Enforcement Agencies (LEAs), e.g., to dismantle their organisation. However, poor data quality and the robustness of criminal networks make effective intervention extremely challenging. Although there exists a large body of work building and applying network scientific tools to green intervene criminal networks, these work often neglect the problems of data incompleteness and inaccuracy. Moreover, there is thus far no comprehensive understanding of the impacts of data quality on the downstream effectiveness of interventions. This work investigates the relationship between data quality and intervention effectiveness based on classical graph theoretic and machine learning-based targeting approaches. Decentralization emerges as a major factor in network robustness, particularly under conditions of incomplete data, which renders intervention strategies largely ineffective. Moreover, the robustness of centralized networks can be boosted using simple heuristics, making targeted intervention more infeasible. Consequently, we advocate for a more cautious application of network science in disrupting criminal networks, the continuous development of an interoperable intelligence ecosystem, and the creation of novel network inference techniques to address data quality challenges.

Keywords: Criminal network; Data quality; Complex system; Organized crime

1 Introduction

Criminal organizations are ubiquitous and the *Dark Networks* that support their operations are threats to our democracy, the rule of law and public safety [1]. Criminal networks operate outside of the law in various contexts, such as drug trafficking rings [2] and terrorist organizations [3]. For example, in 2008, 2.3% of the Australian population whose age are over 14 had consumed methamphetamine within 12 months [2], revealing an underlying public health issue across the country. Terrorist organizations, such as the Global Salafi Jihad (GSJ) network, which includes al-Qaeda and was responsible for large-scale attacks like 9/11, are among the more extensively studied criminal networks. More recently, within the European Union (EU), around thirty transnational criminal networks are active across most member countries, driving violent and exploitative crimes such as burglary and sex trafficking [4].

Although criminal networks are widely recognized and monitored by governments worldwide, disrupting them remains a significant challenge for Law Enforcement Agencies (LEAs) and intelligence agencies. One key obstacle is the advancement of secure communication technologies, which enable criminal organizations to coordinate illegal activities with greater efficiency and reduced detection [5]. The persistence and robustness of these organizations are often reinforced by their adaptability, reliance on corruption, and use of forensic countermeasures, such as encrypted software like SkyECC, to evade monitoring. Over the past two decades, these covert networks have also become increasingly decentralized [6]. Additionally, the collection, management, and interpretation of criminal data are frequently flawed. For large networks, such as the Sicilian Mafia or outlaw motorcycle gangs (OMCGs), surveillance often fails to capture critical communications, resulting in substantial data gaps [7]. Conversely, in smaller networks, intelligence gathered from reports or investigations may be overlooked or lost due to corruption or poor judgment [8]. As a result, data on covert networks is often incomplete, inaccurate, and unreliable. Finally, the diversity of network topologies and dynamics across various types of covert organizations complicates efforts to develop a unified approach to network intervention. Although 'intervention' generally refers to various strategies employed by LEAs to disrupt criminal organizations, we use this term in this paper to specifically refer to targeted methods aimed at influencing individuals in criminal organizations (see Sect. 4.2).

Data quality and the adaptability of criminal organizations are key themes of contemporary research on covert networks and the intervention of these networks. These challenges reflect the robust nature of criminal organizations - that is the ability to evade detection and recover from LEAs' interventions (e.g., arrest, incarceration) [9]. In line with the body of criminal network intervention literature [10–17], we will use network robustness throughout this Article to represent such resilience against interventions. Over the past two decades, network science has emerged as a critical tool for intervening in criminal networks. By leveraging new data sources, such as cellphone call records, network science has been applied to develop interactive strategies that assist in suspect identification and in revealing the hidden structures of criminal organizations [18]. A key approach in these studies is to target and remove individuals by individuals' importance, quantified through centrality measures. Indeed, a substantial body of literature on criminal network intervention evaluates the most effective node-ranking strategies for targeting or apprehending actors within these networks [10, 14, 17, 19].

However, a primary limitation of these studies is their reliance on the assumption of data completeness and accuracy, which can lead to overfitting on potentially flawed data. Not only is data collection extremely challenging in this context [20], but criminal networks can also manipulate their structures—using tactics such as the Remove-One-Attach-Many (ROAM) heuristic (see Sect. 4.3) to obscure leaders and captains [21] or optimizing network design to create nodes with identical centralities [22]. Criminal networks exhibit highly dynamic structures driven by evolutionary processes such as member recruitment, incarcerations, deaths, and the adoption of novel communication channels. For instance, [23] found that eleven Islamic terrorist networks tend to become increasingly centralised and transitive in preparation for specific tasks, such as terrorist attacks. This transformation involves shifting from decentralised, all-channel structures toward networks featuring distinct central hubs, suggesting strategic centralisation orchestrated by criminal leadership rather than organic network evolution. For example, Krebs' terrorist networks typically exhibit high average degree centrality, as do certain criminal enterprises such as the Caviar network, which revolves around a single leader. Similarly, studies of multiple American and British street gangs [24, 25] have identified comparable average centralities to those observed in the Caviar network.

In contrast, a notably different topological trajectory was observed within a New South Wales drug trafficking network, where average degree centrality and network density decreased as the network expanded over time, despite initially presenting centrality comparable to early-stage terrorist networks [26]. These contrasting observations highlight the substantial variability in the topological evolution of criminal networks, driven by their differing operational objectives - and potentially different data collection procedures. Currently, however, there is a lack of systematic understanding concerning how local- and global-level network metrics—including centralisation, density, clustering, and average path distance—relate to the effectiveness of intervention strategies, particularly in scenarios complicated by data incompleteness and inaccuracies [15, 16]. Furthermore, while it is recognised that data incompleteness exacerbates errors in estimating network properties and characteristics [27], little research has specifically investigated when and how these data limitations negatively impact the success of network intervention strategies.

In recent years, increasingly sophisticated intervention strategies have been developed to identify key individuals who significantly influence information diffusion within networks. These techniques commonly involve algorithms leveraging spectral properties, message-passing (belief propagation), and machine learning methodologies to devise optimal intervention strategies [28]. A prominent example is the Collective Influence (CI) algorithm, which aims to minimise the largest eigenvalue of the non-backtracking matrix by calculating a CI score for each node [29]. Subsequently, [30] and [31] introduced machine learning-based node ranking approaches using reinforcement learning and geometric deep learning to determine optimal targets for intervention. These advanced models demonstrated superior performance compared to established algorithms such as CI, Explosive Immunisation (EI, [32]), and Core High-Degree (CoreHD, [33]). However, despite this variety of innovative approaches, it remains unclear whether these newer methods consistently outperform classical intervention techniques across different datasets and network topologies, especially in the presence of data incompleteness and inaccuracies.

Overall, findings from previous studies on the effectiveness of intervention strategies should be interpreted with careful consideration of issues related to missing and inaccurate data. While recent work has begun addressing the impact of missing data in criminal networks (see [15]), these studies have largely concentrated on (1) large datasets and (2) network estimation errors. In this context, the present work examines the core question of how data quality—whether compromised by incompleteness, inaccuracy, or intentional network self-alteration—affects the effectiveness of downstream network interventions in smaller networks. Through a series of percolation experiments, our results indicate that missing data renders most node-ranking methods ineffective at reducing the size of the Largest Connected Component (LCC) in both centralized and decentralized networks. This limitation is further exacerbated by the potential for network topology to be restructured through simple heuristics. Based on these findings, we advocate for heightened awareness of the limitations of network science in disrupting criminal networks under various scenarios of poor data quality and emphasize the need for ongoing advancements in data collection and annotation methods.

Table 1 Data description of the networks being investigated. *N* and *M* indicate the number and edges of the networks respectively. The synthetic networks are generated with approximately constant density to control for the network connectivity for comparative purposes

Type	*N*	*M*	Clustering	Average Distance	Density	Degree centralization
Synthetic						
ER	50	149	0.09	2.36	0.12	0.11
BAHK	50	140	0.52	2.32	0.11	0.39
BA	50	141	0.27	2.26	0.12	0.37
WS	50	150	0.25	2.43	0.12	0.06
Superlinear DN	50	137	0.48	3.29	0.11	0.20
Empirical						
London Gangs	54	315	0.63	2.05	0.22	0.26
New York Cocaine Trafficking Network	28	40	0.34	2.07	0.11	0.84
Ndrangheta mob network	139	1470	0.81	2.33	0.15	0.37
Madrid Bombing Terrorist Network	17	63	0.90	1.59	0.46	0.54

2 Results

While numerous robustness measures exist [34], we choose sequential node percolation for our analysis due to its two primary advantages. First, it is widely used as a method to study network vulnerability. Second, node percolation serves as an abstract representation of real-life interventions in criminal networks, such as the arrest and incarceration of key individuals. For this purpose, we selected various node-ranking methods based on classical centrality measures, as well as more advanced heuristics-based and machine learning-based approaches (see details in Sect. 4). We then used these ranking strategies to conduct percolation experiments on the networks.

We use publicly available static network data, specifically: (1) the London juvenile gang network [35], (2) the 'Ndrangheta network [36], (3) the New York cocaine trafficking ring [37], and (4) the Madrid train bombing terrorist networks [38] (see details of networks studied in Table 1). These networks were selected to provide diversity in size, topology, and organizational goals, allowing for a comparative examination of intervention effectiveness (see Sect. 4.1). Although the data sources occasionally include demographic variables and edge attributes (e.g., relationships between actors), our analysis treats these networks as unattributed, unweighted, and undirected graphs to facilitate more granular manipulation of network topologies. Building on experiments with the baseline networks (i.e., networks as represented in the original datasets (see Sect. 2.1), we conducted simulations to assess intervention effectiveness on perturbed networks (i.e., networks with missing, inaccurate, or topologically altered data).

2.1 Baseline performance of network intervention

In Sect. 2.1, we report the baseline performance of criminal network intervention. We quantify the network intervention effectiveness by approximating Area Under the Curve (AUC) of the size reduction of the Largest Connected Component (LCC) throughout the sequential percolation estimated by the Trapezoidal Rule.

The lower the AUC, the more effective an intervention is. We observe that there exists a generally wide standard deviation of effectiveness σ across all empirical datasets, but notably more so in the earlier phase of the New York cocaine trafficking ring and in most parts in the 'Ndrangheta network. However, the average AUC is much lower for the

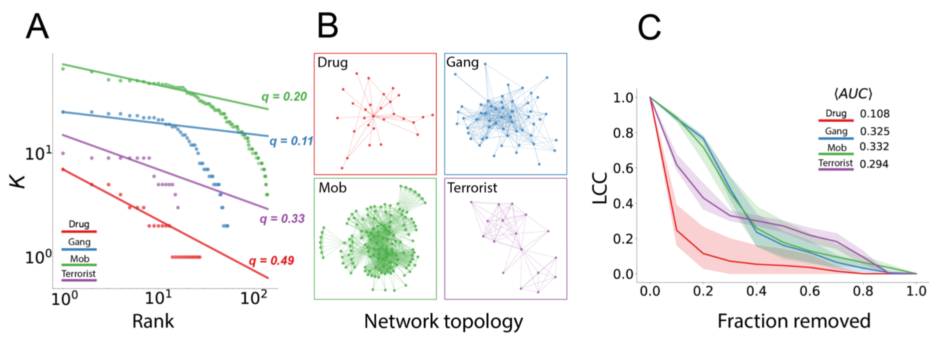

Figure 1 Criminal networks investigated in this work. *A*: Rank-Degree distribution fitted with Rank-Size scaling law $P_k = P_1 k^{-q}$, where P_1 denotes the highest degree and $k \in [1, 2, \ldots, N]$ refers to the rank. Rank is normalized for comparative purposes. Note that not all networks were well-fitted due to varying levels of network centralization. *B*: Visualization of the networks. *C*: Baseline percolation with shaded area is the standard deviation σ and $\langle AUC \rangle$ is the average Area Under the Curve (AUC) of the LCC trajectories across all node-ranking methods

cocaine trafficking network ($\langle AUC \rangle = 0.108$) than the mob network ($\langle AUC \rangle = 0.332$), indicating the resilient nature of the mob network (see Fig. 1).

Another intriguing property is network centralization. As mentioned earlier, decentralization has been shown to be a significant factor for network robustness. To quantify centralization, we used Freeman degree centralization coefficient (C_d) [39]

$$C_d = \frac{\sum_v \max_w c_w - c_v}{n^2 - 3n + 2}, \tag{1}$$

where i denotes the node with highest degree centrality. The denominator $n^2 - 3n + 2$ is the theoretical maximum sum of difference in degree given that a graph with one dominant node (e.g., a star graph) must have a degree of $n - 1$ if self-loops are prohibited. For the other nodes in the graph, then, the degree will automatically be 1 and thus the difference between the dominant node and any follower node is $(n - 1) - 1 = n - 2$. Altogether, the maximum sum of difference for all $n - 1$ dominant-follower pairs is $(n - 2)(n - 1) = n^2 - 3n + 2$.

Our numerical experiments indicate that networks with lower degree centralization—i.e., decentralized networks—generally exhibit greater robustness to network interventions. Notably, degree centralization is closely correlated with the rank-degree distribution. Networks with a rank-degree distribution that closely follows a typical Zipf scaling law, such as the Cocaine trafficking ring ($q = 1.20$) and the Madrid bombing terrorist network ($q = 0.38$), showed lower average AUC values compared to the more resilient, decentralized networks. It is worthwhile to mention that the decentralized networks also exhibits weaker power-law scaling, with $q = 0.27$ and $q = 0.32$ for the gang and mob networks respectively. This result was also consistent in synthetic networks with matching centralization configurations.

2.2 Influence of missing data on intervention effectiveness

In this section, we examine the impact of missing data on network intervention effectiveness. Building on the baseline established earlier, we repeated the percolation experiments using randomly sampled subgraphs of the original graphs, assuming these graphs represent the true underlying network structure of the target organization. To simulate data incompleteness, subgraphs were generated by selecting a random fraction q of nodes from the underlying graph. Each simulation was conducted 10^3 times per node-ranking method across all data completeness scenarios.

We observed that data incompleteness affected different networks similarly, regardless of the extent of incompleteness. For instance, although the New York cocaine trafficking ring and the London gang network had markedly different baseline performance, AUC values increased almost linearly as data completeness decreased. Thus, data incompleteness poses a challenge for the LEAs' irrespective of network topology, although centralized networks—like the New York cocaine trafficking ring and the Madrid bombing terrorist network—were still found to be more vulnerable to intervention under missing data conditions. Sect. 2.2 illustrates the specific effects of data incompleteness on the Ndrangheta network. In general, for more resilient networks, such as the London gang and Ndrangheta networks, data incompleteness proves highly problematic for LEAs, as percolation effectiveness remains weak even with a relatively low percentage of missing data when using previously effective methods like FINDER and CI (see Fig. 2C). In other words, to dismantle these networks effectively, the observer would need near-complete knowledge of the network structure.

We conducted additional quantile regression analysis with network completeness q as the independent variable and intervention efficiency (AUC) as the outcome variable to further confirm their relationship. For instance, on the 'Ndrangheta network, as illustrated in Fig. 2B, we observed a significant negative effect of network completeness on the value

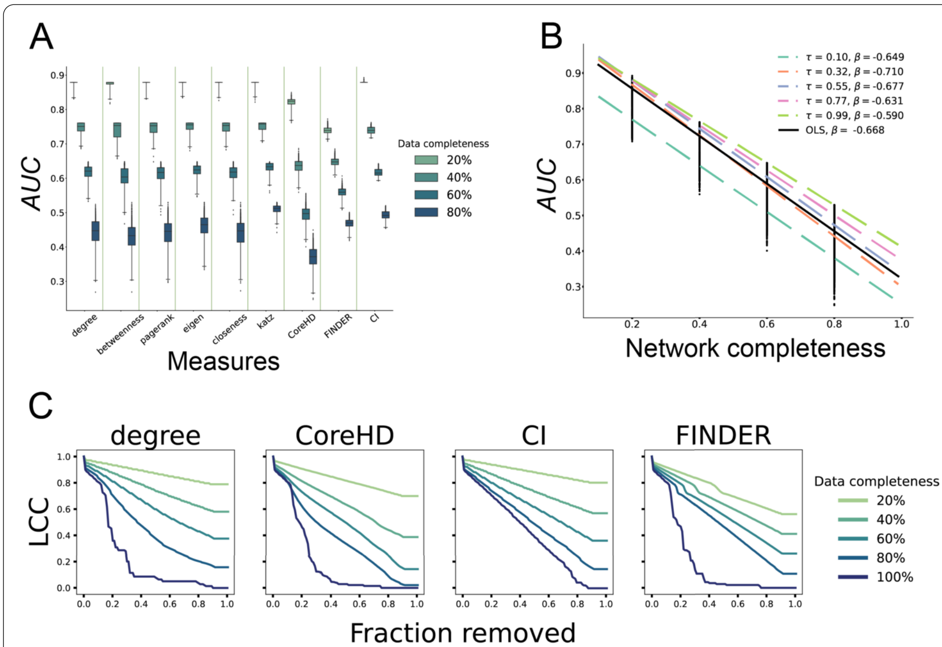

Figure 2 Impact of data incompleteness on percolation effectiveness in the 'Ndrangheta network with 10^3 simulations per node-ranking method across all data incompleteness scenarios. *A*: Boxplots of all node-ranking methods under different data completeness scenarios. Whiskers show the inter-quantile range (IQR) of the AUC and the outliers are indicated by small dots. *B*: Quantile regression of the effect of data completeness on AUC. *C*: Percolation plots of four different node-ranking methods under different data scenarios averaged over simulations

of AUC ($p < .001$, $t = -1672.5$) with a 95% confidence interval of $[-0.686, -0.685]$. This effect, however, varied across quantiles (τ) of the outcome distribution. Specifically, we detected a non-linear association between τ and the regression coefficient β_{q}, showing the strongest negative effect near the median of the efficiency distribution, a weaker effect at the lower end of efficiency (where AUC is high), and the weakest effect at the higher end (where AUC is low).

Similar results were observed in the synthetic networks. Additionally, we found that ER and WS graphs exhibit high robustness against intervention, particularly in scenarios with significant data incompleteness. This robustness makes ER and WS graphs especially challenging to intervene when data incompleteness is present.

2.3 Influence of inaccurate data on intervention effectiveness

We present here the impact of data inaccuracy on intervention effectiveness. Unlike data incompleteness, where nodes may be missing, data inaccuracy retains the same number of observed nodes ($q = 1$). To model data inaccuracy, 10^3 randomly modified graphs G' per inaccuracy scenario were generated based on the inaccuracy rate $p^k \in [0,1]$, representing the proportion of nodes p with k edges that were added, deleted or rewired (non-degree-preserving) per node (see Algorithm 1). For each randomly modified graph, we again performed percolation as outlined in Sect. 2.2, and measured the effectiveness of the intervention methods by computing the AUC of the reduction of the size of LCC. Note that for any node with a number of edge(s) smaller than or equal to k, we do edge modification for all of its edges. In cases where a node has zero degree after modification, the node would be ranked the lowest by all metrics and was removed last.

Algorithm 1 Data Inaccuracy Model

Input: Graph $G = (V, E)$, inaccuracy rate $p^k \in [0,1]$
Output: Modified graph $G' = (V, E')$ with inaccuracy rate p^k
Randomly sample a subgraph $G_{sub} \subseteq G$ with a sampling rate of p
for each node $u \in V(G_{sub})$ **do**
  Randomly select k edges to modify according to the predetermined data inaccuracy scenario for node u
  if Data inaccuracy scenario is edge removal **then**
    if $k \geq \mathrm{Degree}(u)$ **then**
      Remove all edges connected to node u
      Node u remains in G_{sub} with $\mathrm{Degree}(u) = 0$
    else
      Remove k random edges from node u
    end if
  else if Data inaccuracy scenario is edge addition **then**
    Add k random edges to node u
  else if Data inaccuracy scenario is edge rewiring **then**
    Rewire edges of node u with k other random nodes
  end if
end for

The results indicate only marginal differences in intervention effectiveness for graphs with small perturbations through such edge modifications. In the conservative scenario where only one inaccurate link ($k = 1$) can be modified for each node in their ego network, data inaccuracies exerted minimal negative impact on percolation effectiveness. Although a higher level of inaccuracy did delay the collapse of the largest connected component (LCC), shifting the trajectories rightward, this effect was consistent across all types of edge modifications. The influence of different types of edge modification on percolation effectiveness remained qualitatively similar. Consistent with our earlier findings, inaccuracies produced comparable effects across graphs with varying robustness levels. Our comparative analysis of ER and BAHK graphs under data inaccuracies revealed a relatively uniform delay in LCC dismantling with increasing inaccuracy. However, when we consider the less ideal case where multiple edges may be added to each inaccurate node, distinct patterns emerged. For instance, the less robust New York cocaine trafficking network experienced a more pronounced robustness shift due to inaccuracies than the London gang network. Despite general delays in both networks, the robustness shift for the New York network was significantly greater ($\Delta AUC_{p^2=0} \approx 0.01 \rightarrow \Delta AUC_{p^2=1} \approx 0.07$) compared to that of the London network ($\Delta AUC_{p^2=0} \approx 0.223 \rightarrow \Delta AUC_{p^2=1} \approx 0.229$). Additionally, our findings reveal that random data inaccuracy does not always translate to lower percolation effectiveness; bigger inaccuracies do not necessarily correlate with worse outcomes.

2.4 Robustness boosting and intervention effectiveness

Given the understanding that high centralization reduces network robustness against interventions, an important question arises: can we enhance network robustness through leader-hiding techniques? In this section, we present the results of numerical experiments aimed at increasing robustness through topological alterations using the ROAM heuristic.

Figure 3A shows the results of the percolation experiment on a centralized network (i.e., cocaine) and a decentralized network (i.e., mob) after applying the ROAM heuristic. ROAM starts to be very effective in strengthening a centralized network when b and $exec_n$ are sufficiently large. Particularly, when $b \geq 6$ and $exec_n \geq 8$, the effectiveness of the intervention strategies are significantly reduced by a great margin ($0.003 \leq \Delta\text{AUC} \leq 0.15$). Intriguingly, we found that ROAM worked significantly better for the originally centralized networks (i.e., New York cocaine trafficking ring and Madrid bombing network), and was much less effective for networks that are already decentralized. While the ROAM-altered centralized networks did not achieve the same level of robustness as the reference baseline (the WS graph), it is reasonable to conclude that ROAM serves not only as a leader-hiding technique to evade detection but also as an effective robustness-enhancing method with a significant impact on the intervention phase of network intervention.

As a next step, we evaluate how the parameterization of the algorithm may affect the effectiveness of such hiding techniques. The parameters quantify the costs required to run the algorithm in real-life. Particularly, the budget (b) and number of executions ($exec_n$) parameters in ROAM were inspected. Figure 3B illustrates the gain in robustness as measured by the AUC of various LCC trajectories relative to the original graph in ROAM. Assuming that any modification to edges incurs the same cost, increasing the number of executions almost always yielded better payoff when controlling for the total number of edges that can be modified. For example, under the scenario where the total number of edges modified is 24 (e.g., $b = 4$, $exec_n = 6$; $b = 6$, $exec_n = 4$), the payoff is higher when the

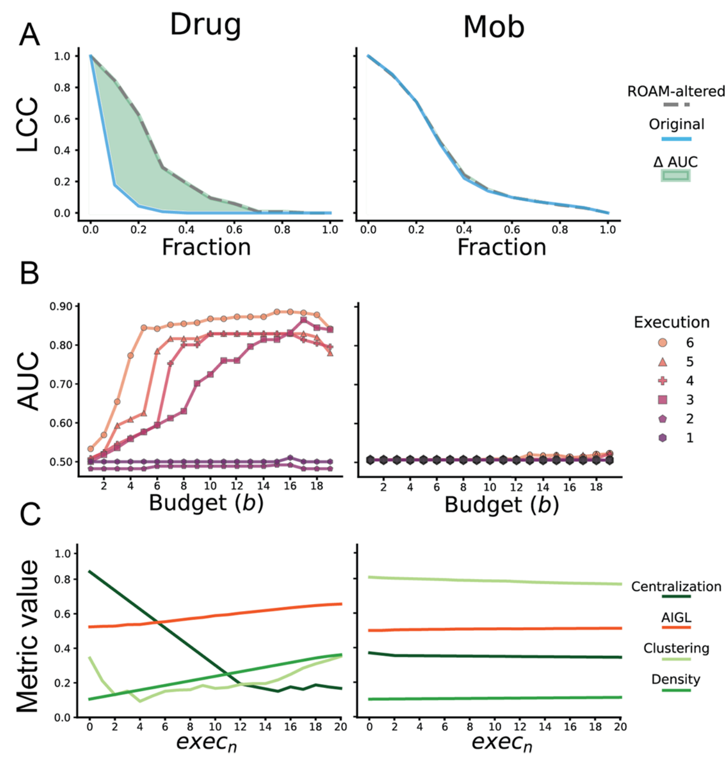

Figure 3 Results of the percolation experiments on the ROAM-altered cocaine trafficking ring and the 'Ndrangheta network ($b = 6$, $exec_n = 8$). *A*: Evolution of LCC under ROAM-altered network (blue line) against the original network (gray dotted line). Green area indicates the positive difference of AUC between the two trajectories. The trajectories are averaging over all measures as intervention strategies. *B*: Change in AUC with different values of *b* and $exec_n$ *C*: Change in the four network statistics over varying $exec_n$ of ROAM

number of execution is higher than the budget. Therefore, perhaps unsurprisingly, there exists a clear relation effect between budget and number of executions of the algorithm. Nonetheless, it should be emphasized that having a low number of executions inhibits the positive influence of budget on the robustness improvement. A general observation is that having a lower number of executions delays the robustness gain even given the large amount of budget devoted to each execution. Particularly, when $1 \leq exec_n \leq 2$, the network robustness does not improve at all. One significant note to point out is that even with higher number of executions and budget, the positive effects of ROAM on AUC for the originally decentralized networks become negligible.

Finally, to understand how ROAM induces structural changes, we conducted additional analyses on the evolution of several global graph properties, namely (1) centralization, (2) average inverse geodesic length (AIGL), (3) average clustering and (4) network density. Figure 3C considers the New York cocaine trafficking ring and the 'Ndrangheta network. Beyond the two representative networks shown, we observe that across all originally centralized networks, for the networks that gained the starkest increase in AUC, (1) centralization has a negative trend and (2) AIGL, (3) clustering and (4) density have positive trend over the iterations. In other words, via ROAM, the centralized networks are in fact in a process of decentralizing and fostering previously distant connections as observed from the increasing clustering coefficient, AIGL and network density. Nonetheless, echoing our findings earlier in this section, originally decentralized graphs did not experience any visible changes in the structure of the graph (see Mob in Fig. 3C).

3 Discussion and conclusion

Data incompleteness and inaccuracy pose significant challenges to law enforcement agencies (LEAs) in effectively disrupting criminal networks [15, 16]. This article aims to explore critical issues in criminal network intervention through the application of network and information-theoretic tools. Our findings reveal that even a 20% level of missing data can severely impair the effectiveness of leading node-ranking methods. In contrast, data inaccuracy only compromises these methods when a high percentage of actors' information is inaccurately captured. This result highlights that data incompleteness extends beyond simply skewing network statistics; it directly undermines the success of network intervention strategies. Specifically, we observed that data incompleteness (i.e., networks with missing nodes) is a more challenging task than data inaccuracy (i.e., networks with modified edges).

Data incompleteness is a significant threat to both the intelligence and implementation phases of network intervention missions. How, then, can we address this challenge? A critical first step towards a more resilient criminal network intervention strategy is investing in consistent, interoperable data infrastructures for effective data sharing (see Sect. 3). This goes beyond simply merging as many data sources as possible, as [20] suggests; the issue often stems from technical inconsistencies across varied data sources. To maximize both the scope and accuracy of network data, it is essential to synthesize "large volumes of disparate data" [40] (p.3) collected from diverse intelligence channels. A recent model of such an integrated security system is the EU interoperability regime established under Regulation (EU) 2019/818, which facilitates police and judicial cooperation via systems like the European Criminal Records Information System (ECRIS), the Europol system, and the Prüm II framework [41]. Prüm II, in particular, aims to "improve, facilitate and accelerate data exchange" by enabling more open sharing of biometric and criminal records across EU member states [42]. However, these regulations and systems often lack the formalization necessary for practical network data representation. Interoperability efforts tend to focus on data storage and exchange (e.g., using the universal message format, UMF), yet they neglect crucial aspects of data fusion, such as defining legitimate criminal contacts (e.g., is co-arrest sufficient to establish a link?) and standardizing the labeling of individuals and their relationships. Consequently, constructing a reliable representation of criminal networks remains a challenge for LEAs. While data-sharing infrastructure continues to evolve, advancements in data collection technology must be matched by collaboration with civil society organizations (e.g., Tech Against Terrorism) and networked investigators (e.g., Bellingcat) to foster a more comprehensive intelligence-gathering process. For example, integrating intelligence sources and employing network analysis can identify critical targets in financial networks to combat money laundering effectively. Crucially, fostering collaboration among law enforcement agencies (LEAs), researchers, and civil society is essential for achieving network disruption objectives set by practitioners. An illustrative example is the COMCRIM project, a Dutch consortium comprising police forces, banks (e.g., ABN AMRO, Rabobank), NGOs (e.g., FAIRWORK), government entities (e.g., the Ministry of Foreign Affairs), and universities, united in their efforts against money laundering [43]. All in all, we encourage the further development of interoperable

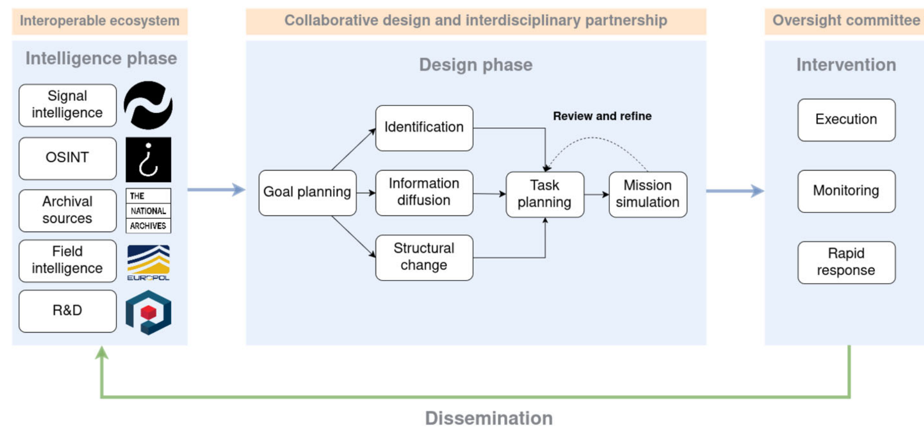

Figure 4 Effective criminal network intervention framework. *Intelligence phase:* Developing interoperable intelligence ecosystem is key to establish efficient data sharing between sources and agencies to improve data accuracy. *Design phase:* Identifying policy goals and designing appropriate mission simulations to ensure effectiveness of intervention. A case in point is a numerical experiment as outlined in this work. *Intervention phase:* Carrying out planned tasks and monitoring the progress of intervention. The results of the intervention should always be fed back to the intelligence phase for further refinement

ecosystems and framework to support data fusion from various intelligence sources (see Fig. 4). Moreover, we urge more collaborative design and interdisciplinary partnerships to facilitate policy goal planning and mission simulations. Finally, when the intervention is to occur, an oversight committee must be present to monitor not only the effectiveness of the intervention, but also the law-abiding and ethical implementations of these interventions (see Sect. 4.2).

Our experiments also reveal that centralized networks are consistently more vulnerable to intervention, though they become more resilient as data quality declines. This finding, combined with highly effective robustness-boosting techniques, is particularly troubling given the trend toward increased decentralization—both technically and socially—in illicit networks over time [44]. Notably, the ROAM leader-hiding technique does more than simply obscure key actors; it significantly enhances the robustness of criminal organizations against network intervention, even in relatively centralized networks. This aligns with expectations, as such techniques effectively reduce degree centralization, thus blurring the lines between centralized and decentralized networks. Consistent with prior research indicating that centralized networks are susceptible to degree-based or value chain-based intervention [20], our results suggest a similar trend: more decentralized networks—both synthetic (e.g., WS, ER) and empirical (e.g., London gangs, 'Ndrangheta networks)—tend to be more robust against non-random intervention strategies across all evaluation metrics. If criminal networks continue to decentralize in response to technological advances, the detrimental effects of data incompleteness may be even greater than previously estimated. Furthermore, as this paper focuses on dismantling smaller sub-units of organized crime networks, large-scale dismantling of poly-criminal networks might exhibit different dynamics. According to [4], approximately 20% of high-risk criminal networks are poly-criminal, meaning they encompass diverse, topologically varied components. With decentralization and poly-criminality on the rise, effectively disrupting these networks remains a critical challenge for LEAs, even with advanced tactics. For these reasons, alongside developing interoperable intelligence systems, we urge the scientific community to advance rigorous network inference techniques—such as network reconstruction and deep learning approaches—that can accurately integrate information from diverse sources.

In summary, future research could extend this study by examining larger networks, such as cryptocurrency transaction networks and illicit peer-to-peer communication systems, to assess the replicability of our findings. We also encourage the exploration of alternative percolation methods, such as triadic or community-based approaches, which may offer insights into intervention and its dynamics in larger and more complex networks. Additionally, percolation experiments on networks with adaptive features—such as recruitment or temporary incarceration modeled through Susceptible-Infected-Recovered (SIR) frameworks—could reveal further realistic impacts of data quality on the effectiveness of intervention strategies.

4 Materials and method

4.1 Data description

The empirical networks were primarily collected through evidence presented in courts with edges between individuals representing an incidence of communication, including but not limited to tapped phone calls, mob conferences and co-appearance in arrests. To validate our result, we generated various unweighted, undirected graphs with several generative models, namely (1) Erdős–Rényi model (ER); scale-free networks with (2) Barabási-Albert model (BA) [45] and (3) Holme and Kim's variation of the BA model (BAHK) [46]; (4) small-world network with Watts-Strogatz model (WS) [47]; (5) superlinear densifying network (SDN) [48]. The models were parameterized to minimize the difference in network density and number of edges to control for the connectivity in the graphs.

4.2 Intervention strategies

The following sections will detail the mathematical rationale of the node ranking tactics. Fig. 5 shows the rank-biased correlation between the ranking tactics. Note that we do not consider bond percolation in this work as site percolation is considerably more efficient in graph dismantling tasks [19].

4.2.1 Heuristics-based intervention

Other than the classical centrality measures (see Table 2), we also used two heuristics-based method for targeting. CI is a heuristic that search for the minimal set of influencers to be targeted to reduce their influence in a network as seen in a typical influence maximization problem [29]. CI index of a node given by

$$\mathrm{CI}(v_i) = (k_i - 1) \sum_{j \in \partial B(i,\ell)} (k_j - 1), \tag{2}$$

where k_i is the degree of node v_i, $B(i,\ell)$ is the ball centring on node v_i and $\partial B(i,\ell)$ are the nodes at the frontiers of the ball. In simple terms, CI of a length $\ell = 2$ is the product between the sum of the degree of all nodes located at the shortest path distance exactly at 2 from node v_i and the degree of the node v_i itself. This quantity is a scalable method to search for minimal sets of nodes [29], which is extremely useful for finding influential players in very large complex networks. [49] developed an even more efficient computational method using max-heap, with the computational complexity $O(n \log n)$.

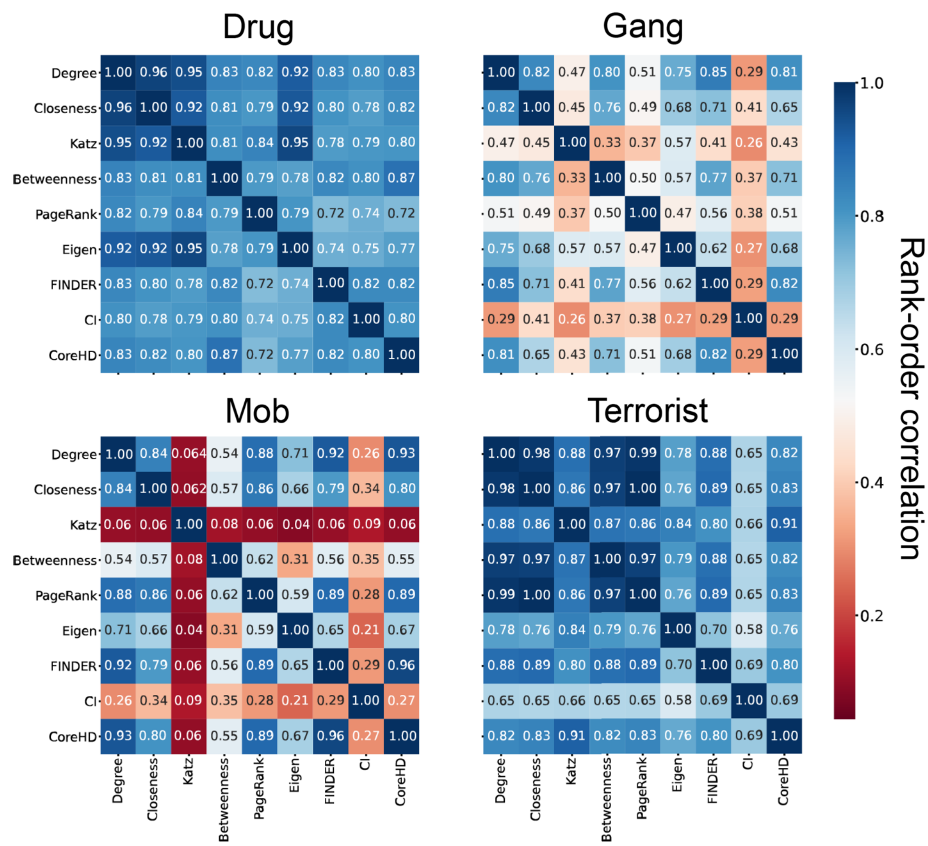

Figure 5 Rank-biased overlap similarity of node-ranking methods used in the percolation experiment. Higher values indicate higher similarity between node-ranking methods

Table 2 Node-ranking methods used in this work. Computational complexity is assumed optimal for sparse networks. Note that for eigenvector and PageRank, k indicates the iterations needed for convergence

Category	Method	Computational Complexity
Centrality-based	Degree centrality	$O(m)$
	Eigenvector centrality	$O(km)$
	Katz centrality	$O(n^3)$
	PageRank centrality	$O(km)$
	Closeness centrality	$O(nm)$
	Betweenness centrality	$O(nm)$
Heuristics-based	Collective Influence (CI)	$O(n \log n)$
	Core High-Degree (CoreHD)	$O(n)$
Machine Learning	FINDER	$O(n + m + n \log n)$

Another heuristic is Core High-Degree (CoreHD), an approach that utilizes degree-based decycling - the disintegration of cycles [33]. CoreHD starts by finding the kcore and obtain the degree of the nodes, then finding the set of nodes V_{HD} with the highest degree in the kcore, given that the size of kcore > 0. If $|V_{\mathrm{HD}}| > 1$, a random node is chosen to be removed. We then update the kcore and the degrees of the nodes until the kcore disappears. Finally, CoreHD will perform tree-breaking and greedy insertion - a procedure to reinsert nodes that were previously unncessarily removed from the decycling process.

This method is similar to the degree centrality-based intervention, but it focuses on adaptive node removal inside the kcore with greedy reinsertion, a technique designed to minimize the number of unnecessary nodes removed during the process of decycling. This method is extremely fast with a computational complexity of $O(n)$ for sparse networks and generally as effective as other message-passing decycling methods such as Min-Sum.

4.2.2 Machine learning-based intervention

We also consider a popular pre-trained Graph Neural Network model (GNN), the so-called FInding key players in Networks through DEep Reinforcement learning (FINDER) [30], in order to target nodes in the network. FINDER is a deep reinforcement learning framework for optimal percolation problems proposed by [30]. Formally, FINDER aims to minimize the accumulated normalized connectivity (ANC)

$$R(v_1, v_2, \dots, v_N) = \frac{1}{N}\sum_{k=1}^{N}\frac{\sigma(G\backslash\{v_1, v_2, \dots, v_k\})}{\sigma(G)}, \tag{3}$$

where N is the number of nodes in G and $v_i \in V$ indicates the ith node to be percolated from the graph, $\sigma(G\backslash\{v_1, v_2, \dots, v_k\})$ is the connectivity of the graph after v_i is removed from the initial graph and, intuitively, $\sigma(G)$ is the connectivity of the original graph. The connectivity metric in FINDER can be any well-defined network metric (i.e., network robustness metrics), making it an extremely adaptable method to consider different types of metrics, such as Von Neumann entropy and spectral gap.

The model contains two phases, namely the offline training phase and the online application phase. In the first phase, synthetic graphs are generated based on different network generative models. These graphs are randomly sampled for the agent to play the game - an episode of a crucial node identification process - where the agent's action is to remove the chosen node. As mentioned, the reward to such an action is defined by the ANC, and the larger the marginal decrease of ANC, the more reward an agent will obtain. The graphs are encoded with tunable parameters Θ_e using inductive graph representational learning to aggregate node embedding vectors as node features to obtain their latent structural position in a graph. After capturing the node embedding, the embedding is then decoded with tunable parameters Θ_d as a scalar Q, a set of scores that assesses the potentials of any given actions. A multilayer perceptron with RELU activation is used to generate the output layer containing the Q values. Using the ϵ-greedy strategy under an exploration-exploitation framework, the action with the highest Q will be adopted with a probability of (1- ϵ), and a random action will be undertaken with a probability of ϵ. ϵ decreases linearly from 1.0 to 0.05 over episodes, symbolizing that a more experienced agent will make decisions based on its past learning (i.e., exploitation) rather than exploring new options in comparison to a less experienced agent. When a game is completed, n-step transitions $(S_i, A_i, R_{(i,i+n)}, S_{i+n})$ are collected. M most recent transitions are then stored in the experience replay buffer, a memory storage technique commonly used Deep-Q learning models. With these memories, the agent is updated with a new set of parameters Θ_e and Θ_d for the encoder and decoding processes respectively. Adam gradient descent updates are used to compute the loss from the randomly sampled experiences from the M most recent memories.

During the application phase, the empirical network will be fed to the model and encoded into a embedding vector with lower dimensions, and then the model will infer the

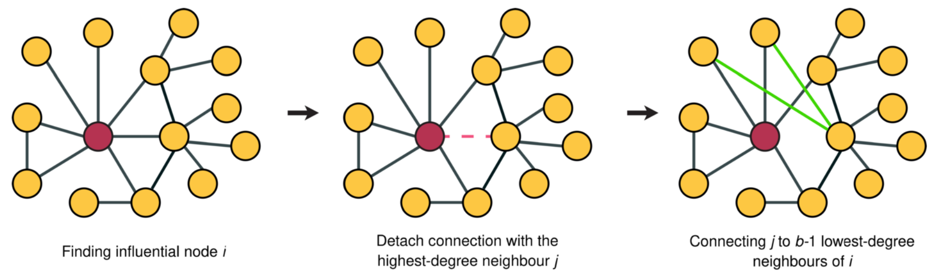

Figure 6 One execution of the Remove-One-Attach-Many heuristic with a budget $b = 3$

Q-value for each node (i.e., the reward when such node is removed from the network) using the decoder. Note that in practice the model selects nodes in batches that can maximize Q instead of the computing Q for each node to reduce computational complexity to $O(|E| + |V| + |V| \log |V|)$. For example, instead of choosing the node v_0, FINDER will evaluate the Q-value of removing the set of $\{v_0, v_1, v_2, v_5\}$.

The model used in this paper is pre-trained by the authors in the original FINDER paper. The model was trained on ER, WS and BA networks ($n_{\text{total}} = 2 \times 10^6$) with each synthetic network containing 30 to 50 nodes. The model sets $M = 5 \times 10^4$ for the experience replay buffer. Because of the batch selection strategy, FINDER only returns the most effective set of nodes to be percolated from the complete graph, meaning that some nodes will be omitted from FINDER. To resolve this problem, the final ordered list of nodes to be removed from graph contains two separate subsets: (1) one with the nodes provided by FINDER and (2) one imputed randomly with the residual nodes in the network. Because subset 1 is not ordered, it is possible that the effectiveness of FINDER under our scenario of sequential node removal is not optimal. However, FINDER has overall been proven to be effective even compared to node rankings that are theoretically designed to be completely ordered.

4.3 Remove-One-Attach-Many heuristic

Finally, we also used a leader-hiding heuristic called Remove-One-Attach-Many (ROAM), an efficient method to hide the leading actor in a network simply by rewiring links of the most central person in a network (see Fig. 6) [21], to boost the robustness of a network. The algorithm takes two parameters, b and $exec_n$, which are the budget available for link addition and removal (see Sect. 6) and the number of consecutive execution of ROAM respectively. In practice, ROAM searches for the evader $v^\dagger$. It then detaches the edge between $v^\dagger$ and its most connected neighbour v_0. By doing so, we reduce the centrality of the leader. Then, to recover the loss of influence of the leader due to reduced connections, we artificially add $b - 1$ links between v_0 and its least connected neighbours. The success of ROAM, then, relies on b as well as how many times we execute the heuristic. In this work, the evader was chosen to be the actor with the lowest combined rank of degree, betweenness and closeness centrality measures, resembling a leader in the network.

Abbreviations

GSJ, Global Salafi Jihad; EU, European Union; LEAs, Law Enforcement Agencies; OMCG, Outlaw motorcycle gangs; ROAM, Remove-One-Attach-Many heuristic; CI, Collective Influence; EI, Explosive Immunisation; CoreHD, Core High-Degree; LCC, Largest Connected Component; AUC, Area Under the Curve; FINDER, Finding key players in Networks through deep reinforcement learning; AIGL, Average inverse geodesic length; ECRIS, European Criminal Records Information System; UMF, Universal Message Format; SIR, Susceptible-Infected-Recovered; ER, Erdős–Rényi model; BA, Barabási-Albert model; BAHK, Holme and Kim's variation of the Barabási-Albert model; WS, Watts-Strogatz model; GNN, Graph Neural Network; ANC, Accumulated normalized connectivity; DQN, Deep-Q Network; SDN, Superlinear densifying network.

Acknowledgements
W.N.Y. thanks Bernie Hogan (Oxford Internet Institute) for his ideas during the conception phase of this work and Rafael Prieto-Curiel (Complexity Science Hub) for his insightful feedback on this work during and after the Conference on Complex Systems '24 in Exeter, United Kingdom. R.L. acknowledges support from the EPSRC grants EP/V013068/1, EP/V03474X/1 and EP/Y028872/1. R.D.C. acknowledges support from the Lagrange Project of the ISI Foundation funded by CRT Foundation.

Author contributions
W.N.Y and R.L. conceived the study, developed the theoretical framework and designed the experiments. W.N.Y. performed all experiments, computation and analyses. R.L. supervised the project. W.N.Y. and R.D.C. produced the figures. All authors discussed the results and contributed to the final manuscript.

Funding information
R.L. acknowledges support from the EPSRC grants EP/V013068/1, EP/V03474X/1 and EP/Y028872/1.

Data Availability
All data is publicly available and detailed in the manuscript. Gang network is from: https://sites.google.com/site/ucinetsoftware/datasets/covert-networks/london-gang Mob network is from: https://comeetie.github.io/greed/reference/Ndrangheta.html Drug network is from: https://sites.google.com/site/ucinetsoftware/datasets/covert-networks/cocaine-dealing-natarajan Terrorist network is from: https://networks.skewed.de/net/train_terrorists

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Complex Connections Lab, Network Science Institute, Northeastern University London, Devon House, London, E1W 1LP, UK. [2]Mathematical Institute, University of Oxford, Andrew Wiles Building, Oxford, OX2 6GG, UK. [3]Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford, OX1 3JS, UK. [4]ISI Foundation, Via Chisola 5, Turin, 10126, Italy.

Received: 6 January 2025 Accepted: 27 April 2025 Published online: 12 May 2025

References

1. Raab J, Milward HB (2003) Dark networks as problems. J Public Adm Res Theory 13(4):413–439. https://doi.org/10.1093/jopart/mug029
2. Bright DA, Hughes CE, Chalmers J (2011) Illuminating dark networks: a social network analysis of an Australian drug trafficking syndicate. Crime Law Soc Change 57(2):151–176. https://doi.org/10.1007/s10611-011-9336-z
3. Xu J, Chen H (2008) The topology of dark networks. Commun ACM 51(10):58–65. https://doi.org/10.1145/1400181.1400198
4. European Union Agency for Law Enforcement Cooperation (2024) Decoding the EU's Most Threatening Criminal Networks. Publications Office, LU. https://data.europa.eu/doi/10.2813/811566. Accessed 2024-07-16
5. Crossley N, Edwards G, Harries E, Stevenson R (2012) Covert social movement networks and the secrecy-efficiency trade off: the case of the uk suffragettes (1906–1914). Soc Netw 34(4):634–644. https://doi.org/10.1016/j.socnet.2012.07.004
6. Cui D, Ou C, Lu X (2024) Evolution of the global terrorist organizational cooperation network. PLoS ONE 19(1):0281615. https://doi.org/10.1371/journal.pone.0281615
7. Baker WE, Faulkner RR (1993) The social organization of conspiracy: illegal networks in the heavy electrical equipment industry. Am Sociol Rev 58(6):837. https://doi.org/10.2307/2095954
8. Morselli C (2009) Inside criminal networks. Springer, Berlin. https://doi.org/10.1007/978-0-387-09526-4
9. Elteren C, Vasconcelos VV, Lees M (2024) Criminal organizations exhibit hysteresis, resilience, and robustness by balancing security and efficiency. https://doi.org/10.48550/ARXIV.2403.03720
10. Duijn PAC, Kashirin V, Sloot PMA (2014) The relative ineffectiveness of criminal network disruption. Sci Rep 4(1). https://doi.org/10.1038/srep04238
11. Ozgul F, Erdem Z (2015) Deciding resilient criminal networks. In: Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015. ASONAM '15. ACM, New York, pp 1368–1372. https://doi.org/10.1145/2808797.2808857
12. Catanese S, De Meo P, Fiumara G (2016) Resilience in criminal networks. Atti Accad Pelorit Pericol Cl Sci Fis Mat Nat 94(2):1–1119. https://doi.org/10.1478/AAPP.942A1

13. Duxbury SW, Haynie DL (2019) Criminal network security: an agent-based approach to evaluating network resilience*. Criminology 57(2):314–342. https://doi.org/10.1111/1745-9125.12203
14. Cavallaro L, Ficara A, De Meo P, Fiumara G, Catanese S, Bagdasar O, Song W, Liotta A (2020) Disrupting resilient criminal networks through data analysis: the case of Sicilian mafia. PLoS ONE 15(8):0236476. https://doi.org/10.1371/journal.pone.0236476
15. De Moor S, Vandeviver C, Vander Beken T (2020) Assessing the missing data problem in criminal network analysis using forensic dna data. Soc Netw 61:99–106. https://doi.org/10.1016/j.socnet.2019.09.003
16. Ficara A, Curreri F, Fiumara G, De Meo P, Liotta A (2022) Covert network construction, disruption, and resilience: a survey. Mathematics 10(16):2929. https://doi.org/10.3390/math10162929
17. Mora Tostado S, Hernández-Vargas EA, Núñez-López M (2024) Modeling human trafficking and the limits of dismantling strategies. Soc Netw Anal Min 14(1). https://doi.org/10.1007/s13278-024-01208-x
18. Zhou P, Liu Y, Zhao M, Lou X (2017) A proof of concept study for criminal network analysis with interactive strategies. Int J Softw Eng Knowl Eng 27(04):623–639. https://doi.org/10.1142/s0218194017500231
19. Holme P, Kim BJ, Yoon CN, Han SK (2002) Attack vulnerability of complex networks. Phys Rev E 65(5). https://doi.org/10.1103/physreve.65.056109
20. Duijn PAC, Sloot PMA (2015) From data to disruption. Digit Investig 15:39–45. https://doi.org/10.1016/j.diin.2015.09.005
21. Waniek M, Michalak TP, Wooldridge MJ, Rahwan T (2018) Hiding individuals and communities in a social network. Nat Hum Behav 2(2):139–147. https://doi.org/10.1038/s41562-017-0290-3
22. Dey P, Medya S (2019) Covert networks: how hard is it to hide? arXiv. https://doi.org/10.48550/ARXIV.1903.05832
23. McMillan C, Felmlee D, Braines D (2019) Dynamic patterns of terrorist networks: efficiency and security in the evolution of eleven Islamic extremist attack networks. J Quant Criminol 36(3):559–581. https://doi.org/10.1007/s10940-019-09426-9
24. Gunnell D, Hilier J, Blakeborough L (2016) Social Network Analysis of an Urban Street Gang Using Police Intelligence Data. Research Report 89. Technical report, Home Office
25. Sierra-Arevalo M, Papachristos AV (2015) Social network analysis and gangs. Wiley, New York. https://doi.org/10.1002/9781118726822.ch9
26. Bright DA, Delaney JJ (2013) Evolution of a drug trafficking network: mapping changes in network structure and function across time. Glob Crime 14(2–3):238–260. https://doi.org/10.1080/17440572.2013.787927
27. Wang DJ, Shi X, McFarland DA, Leskovec J (2012) Measurement error in network data: a re-classification. Soc Netw 34(4):396–409. https://doi.org/10.1016/j.socnet.2012.01.003
28. Artime O, Grassia M, De Domenico M, Gleeson JP, Makse HA, Mangioni G, Perc M, Radicchi F (2024) Robustness and resilience of complex networks. Nat Rev Phys 6(2):114–131. https://doi.org/10.1038/s42254-023-00676-y
29. Morone F, Makse HA (2015) Influence maximization in complex networks through optimal percolation. Nature 524(7563):65–68. https://doi.org/10.1038/nature14604
30. Fan C, Zeng L, Sun Y, Liu Y-Y (2020) Finding key players in complex networks through deep reinforcement learning. Nat Mach Intell 2(6):317–324. https://doi.org/10.1038/s42256-020-0177-2
31. Grassia M, De Domenico M, Mangioni G (2021) Machine learning dismantling and early-warning signals of disintegration in complex systems. Nat Commun 12(1). https://doi.org/10.1038/s41467-021-25485-8
32. Clusella P, Grassberger P, Pérez-Reche FJ, Politi A (2016) Immunization and targeted destruction of networks using explosive percolation. Biophys Rev Lett 117(20). https://doi.org/10.1103/physrevlett.117.208301
33. Zdeborová L, Zhang P, Zhou H-J (2016) Fast and simple decycling and dismantling of networks. Sci Rep 6(1). https://doi.org/10.1038/srep37954
34. Schwarze AC, Jiang J, Wray J, Porter MA (2024) Structural robustness and vulnerability of networks. ArXiv. https://doi.org/10.48550/ARXIV.2409.07498
35. Grund TU, Densley JA (2015) Ethnic homophily and triad closure: mapping internal gang structure using exponential random graph models. J Contemp Crim Justice 31(3):354–370. https://doi.org/10.1177/1043986214553377. Accessed 2024-06-24
36. Legramanti S, Rigon T, Durante D, Dunson DB (2022) Extended stochastic block models with application to criminal networks. Ann Appl Stat 16(4). https://doi.org/10.1214/21-aoas1595
37. Natarajan M (2006) Understanding the structure of a large heroin distribution network: a quantitative analysis of qualitative data. J Quant Criminol 22(2):171–192. https://doi.org/10.1007/s10940-006-9007-x
38. Walther OJ, Christopoulos D (2014) Islamic terrorism and the malian rebellion. Terrorism Polit Violence 27(3):497–519. https://doi.org/10.1080/09546553.2013.809340
39. Freeman LC (1978) Centrality in social networks conceptual clarification. Soc Netw 1(3):215–239. https://doi.org/10.1016/0378-8733(78)90021-7
40. Spyropoulos AZ, Bratsas C, Makris GC, Garoufallou E, Tsiantos V (2023) Interoperability-enhanced knowledge management in law enforcement: an integrated data-driven forensic ontological approach to crime scene analysis. Information 14(11):607. https://doi.org/10.3390/info14110607
41. Giannakoula A, Lima D, Kaiafa-Gbandi M (2020) Combating crime in the digital age: a critical review of EU information systems in the area of freedom, security and justice in the post-interoperability era: challenges for criminal law and personal data protection. Brill Res Perspect Transnatl Crime 2(4):1–97. https://doi.org/10.1163/24680931-12340010. Accessed 2024-07-13
42. European Commission New Prüm II Regulation. https://ec.europa.eu/commission/presscorner/detail/en/ip_23_5870. Accessed 2024-07-13
43. Voorhout JEB (2020) Combatting human trafficking holistically through proactive financial investigations. J Int Crim Justice 18(1):87–106. https://doi.org/10.1093/jicj/mqaa013
44. McCarthy-Jones A, Doyle C, Turner M (2020) From hierarchies to networks: the organizational evolution of the international drug trade. Int J Law Crime Justice 63:100436. https://doi.org/10.1016/j.ijlcj.2020.100436
45. Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74(1):47–97. https://doi.org/10.1103/revmodphys.74.47
46. Holme P, Kim BJ (2002) Growing scale-free networks with tunable clustering. Phys Rev E 65(2). https://doi.org/10.1103/physreve.65.026107

47. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442. https://doi.org/10.1038/30918
48. Lambiotte R, Krapivsky PL, Bhat U, Redner S (2016) Structural transitions in densifying networks. Phys Rev Lett 117(21). https://doi.org/10.1103/physrevlett.117.218301
49. Morone F, Min B, Bo L, Mari R, Makse HA (2016) Collective influence algorithm to find influencers via optimal percolation in massively large social media. Sci Rep 6(1). https://doi.org/10.1038/srep30062

Publisher's note